DOCUMENT RESUME

ED 445 036	TM 031 641

AUTHOR	Li, Yuan H.; Lissitz, Robert W.
TITLE	Applications of the Analytically Derived Asymptotic Standard Errors of IRT Item Parameter Estimates.
PUB DATE	2000-04-00
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Error of Measurement; *Estimation (Mathematics); *Item Response Theory; *Maximum Likelihood Statistics; Prediction; Simulation; Test Items
IDENTIFIERS	Asymptotic Distributions; Partial Credit Model

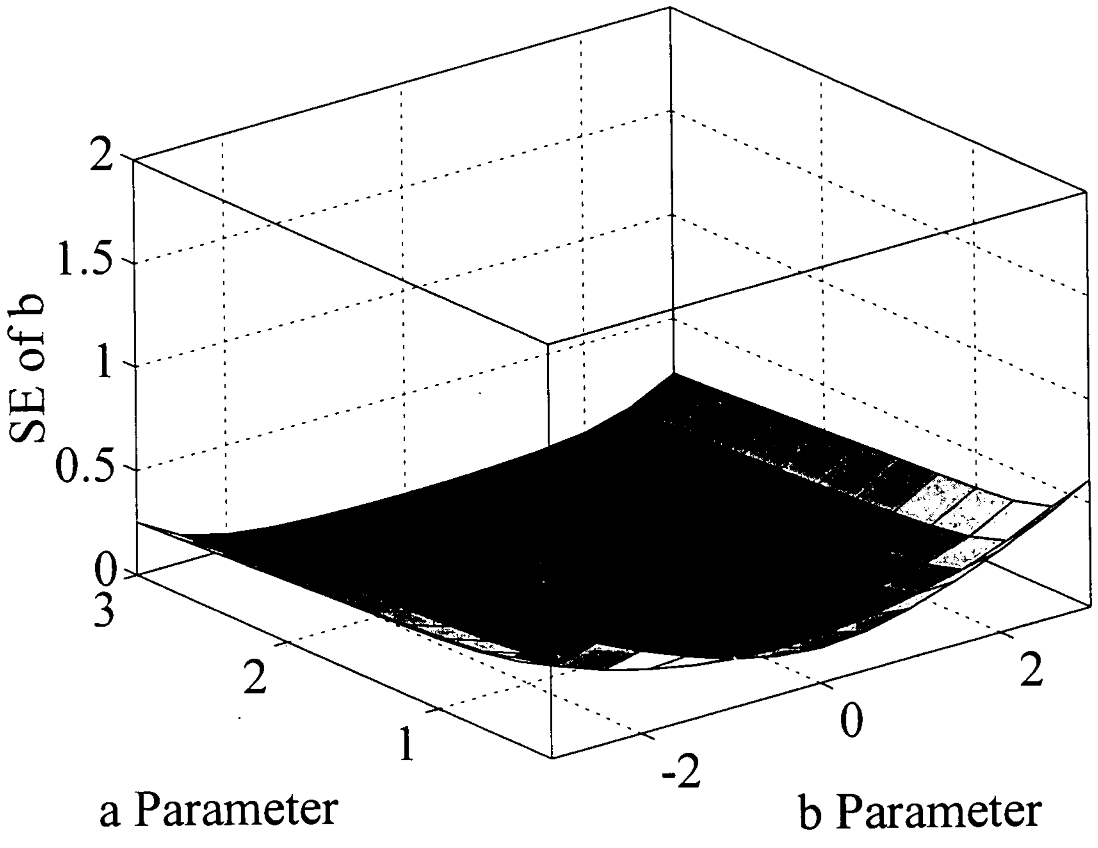
ABSTRACT

The analytically derived expected asymptotic standard errors (SEs) of maximum likelihood (ML) item estimates can be predicted by a mathematical function without examinees' responses to test items. The empirically determined SEs of marginal maximum likelihood estimation/Bayesian item estimates can be obtained when the same set of items is repeatedly estimated from test data. Understanding the consistency of SEs yielded from both approaches is of primary concern for the application of the analytic SEs. In most cases in the simulation conducted, SEs yielded from both approaches were similar, especially for the generalized partial credit model (E. Muraki, 1992). This finding encourages test practitioners and researchers to apply the asymptotic SEs of item estimates to the following applications: (1) practical testing situations; (2) item-linking studies; and (3) predicting the SEs of equating scores for the item response theory (IRT) true-score procedures (Lord, 1982) without examinees' responses to test items. Three dimensional graphical representation of the analytic SEs of item estimates has also been provided for better understanding of several widely used IRT models. (Contains 32 references.) (Author/SLD)

Applications of the Analytically Derived Asymptotic Standard Errors of IRT Item Parameter Estimates

Yuan H. Li
Prince Georges County Public Schools, Maryland

Robert W. Lissitz
University of Maryland at College Park



Paper presented at the annual meeting of the American Educational Research Association
April, 24-28, 2000, New Orleans, LA

Abstract

The analytically derived expected asymptotic standard errors (SEs) of ML (maximum likelihood) item estimates can be predicted by a mathematical function without examinees' responses to test items. The empirically determined SEs of MMLE (marginal maximum likelihood estimation) / Bayesian item estimates can be obtained when the same set of items are repeatedly estimated from test data.

Understanding the consistency of SEs yielded from both approaches is of primary concern for the applications of the analytic SEs. In most cases, the SEs yielded from both approaches were very similar, especially for the generalized partial credit model (Muraki, 1992). This finding encourages test practitioners and researchers to apply the asymptotic SEs of item estimates to the following applications: (a) practical testing situations, (b) item-linking studies, and (c) predicting the SEs of equating scores for the IRT true-score procedure (Lord, 1982) without examinees' responses to test items.

Three-D graphical presentation for the analytic SEs of item estimates has also been provided for better understanding several widely-used IRT models.

Key Index: Asymptotic Standard Errors; Item Parameter Estimates,

Item Response Theory (IRT)

----

*: Paper presented at the annual meeting of the American Educational Research Association April, 24-28, 2000, New Orleans, LA

Applications of the Analytically Derived Asymptotic Standard Errors of IRT Item Parameter Estimates

I. Introduction

A. Background and Motivation

IRT (item response theory) has been widely employed in large-scale testing programs. The successful application of IRT to practical measurement problems relies heavily on the degree of precision of the item- parameter estimates.

The error in the parameter estimates is narrowly defined as the amount of variance around the true parameter value. A variety of legitimate factors can cause errors in the parameter estimates. Four primary factors are often highlighted in the literature. The first one is related to the assumptions made for IRT models that cannot be strictly met, at lease to some extent (Ackerman, 1992), e.g., local independence. The second one is related to the estimation method such as the joint maximum likelihood estimation (JMLE) that may not converge to the true values as the sample size and number of items increase (for detailed discussion, see Baker, 1992). Also, the marginal maximum likelihood estimation (MMLE) may not produce true values if an incorrect prior ability distribution is specified during the process of parameter estimation (Seong, 1990). The third one is associated with model mis-fit (Tam & Li, 1997). For instance, if the dimensionality for item response matrices from a test is multidimensional rather than unidimensional, inaccurate item parameter estimates could occur when a unidimensional IRT model is applied to these item response matrices (Ackerman, 1992; Reckase, 1985). The fourth factor is related to practical limitations. For instance, the sample size used for parameter estimation is not large enough or the examinees' abilities are not heterogeneous so that standard errors (SEs) of item estimates for too hard or too easy items become large (Stocking, 1990).

The magnitude of a SE of an item parameter is an index of the precision of an item estimate. The SE index of an item estimate plays a substantive role in IRT applications. The numerical value SE may not be straightforwardly calculated due to the problems indicated above. However, if the model selected fits test data, the maximum likelihood (ML) estimate is chosen for item calibration, and the examinee's ability distribution is known, an analytic approach to compute the SEs for any set of item parameters and sample size exists (Thissen & Wainer, 1982). No examinees' responses to test items are required. Such mathematical expression for this relationship has been developed by Thissen and Wainer (1982) for unidimensionally dichotomous IRT models and was modified for the unidimensionally polytomous IRT models (Li, Lissitz & Yang, 1999) and for multidimensionally dichotomous IRT models (Li & Lissitz, 2000). When all assumptions used to yield SEs of item estimates are not completely true in practice, the SEs obtained by this analytic approach represent lower limits for actual SEs (Thissen & Wainer, 1982).

The analytically expected asymptotic SEs (called AEA-SE) of item estimates can be used for detecting the potential problems of the application of ML estimation to an IRT model without real test data. For instance, Thissen and Wainer (1982) used a three-D graphical representation to explore the relationship between the expected SEs of item difficulties and the bivariate function of item discrimination and difficulty estimates for the three-parameter model (Three-PL, refer to Lord, 1980). Those plots presented in their study clearly showed that the magnitudes of SEs of the ML difficulty estimates were unacceptably large except when the sample sizes are enormously large (e.g., more than 100,000 examinees). This finding suggested that ML is not an appropriate candidate to be chosen for estimating Three-PL item parameters when the lower asymptote parameter is poorly estimated (Gruijter, 1984). Similarly, this application can be extended to explore other IRT models, such as the generalized partial credit model (GPCM, Muraki, 1992) that has been widely employed in current testing programs, such as the National Assessment of Educational Progress (NAEP, Beaton & Zwick, 1992).

When researchers or test practitioners are interested in a set of item parameters found in literature, in which the corresponding SEs of item estimates were not reported, the analytic approach provides them a sense of how large standard errors of the ML item estimates might be under specific situations. For example, when Li and Lissitz (2000) evaluated how sensitive the three multidimensional IRT (MIRT) item-linking methods (developed in their study) were to the accuracy of item parameter estimates, the analytic approach was used for modeling random errors of the set of MIRT item estimates (Form24b, see Reckase, 1985). The current application facilitates the simulation study of investigating which item-linking methods can tolerate the random (or sampling) errors of item estimates better.

On the other hand, caution should be exercised on the AEA approach to item linking studies. For instance, it could happen that when the Three-PL item linking was conducted, longer tests produced less accurate estimates of item linking coefficients than shorter tests (Li, Lissitz & Yang, 1999). The reasons for this will be discussed later. This unexpected result highlighted the issue of how to appropriately employ the AEA-SEs of item estimates in the context of research studies.

The MMLE/Bayesian estimation, incorporated with the additional information of the priors of item estimates, was often employed in the estimation process, such as with the IRT estimation computer programs, BILOG (Mislevy & Bock, 1990) and PARSCALE (Muraki & Bock, 1996) for dichotomously and polytomously scored items. The empirical SEs and BIASs of MMLE/Bayesian item estimates (called EMB-SEs) can be obtained by the replication approach illustrated in the section of Methodology. In general, as the number of replications increases, the EMB-SE estimate becomes relatively stable and accurate.

The AEA-SE is much less tedious to obtain than the replication approach. When we attempt to apply it, the issue of how consistent the AEA-SEs and EMB-SEs are under similar conditions is critical to test practitioners.

B. Research Purposes

As indicated, without real test data the three-D graphical presentation for AEA-SEs of item estimates has been used for detecting the possible problems of the ML estimation method to the Three-PL model (Thissen & Wainer, 1982). As the GPCM has been increasingly used to model the polytomously scored item responses, the extension of this graphical procedure to the GPCM model will provide test practitioners better understanding of this model.

Understanding the consistency of SEs yielded from the analytic approach and from the replication approach is the first step to successfully applying the asymptotic SEs of item estimates to practical testing situations and in the context of research studies. This issue will be pursued in this study. We expect the results generated from this study to provide valuable knowledge of AEA-SEs, as well as their similarity to the EMB-SEs that can serve as a base for comparison.

Finally, detailed descriptions on how to employ the AEA-SEs on the following applications will be included: (a) practical testing situations, (b) item-linking studies, and (c) predicting the SEs of equating scores for the IRT-true score procedure (Lord, 1982) without examinees' responses to test items.

II. The Analytically Derived Asymptotic Expected SEs of Item Estimates

A. IRT Models

In recent years, we have seen growing use of a test with mixed item formats, e.g., tests used for NAEP. These may consist of multiple-choice and short-response items (dichotomously scored), as well as the constructed-response items (polytomously scored) that often occur when using performance testing.

When test data are collected from mixed-format items, simultaneously fitting different IRT models to different types of items of the same scale (as described by Thissen, 1993) has been employed in several large-scale testing programs (e.g., NAEP). When modeling this sort of test data, the two-parameter model (Two-PL) is often used for handling the item responses generated by the short-response dichotomous items. The Three-PL is used for the multiple-choice dichotomous item responses and the GPCM for the constructed-response (or assay) polytomous responses. The AEA-SEs calculated from these three models were explored, with test data from a mixed-format test with dichotomously and polytomously scored items.

1.Three-PL and Two-PL Logistic Models

The commonly-used Three-PL logistic IRT model was used to model the dichotomous scored items in this study. Under the Three-PL model, the probability, P_{ij}, of a correct response to an item i for an examinee j with ability θ_j is given by (Lord, 1980):

$$P_{ji}(\theta_j) = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i)}{1 + \exp(Da_i(\theta_j - b_i)} \tag{1}$$

where

the symbol of "exp" stands for the mathematical function of the natural logarithm exponential,
a_i is the item discrimination,
b_i is the item difficulty,
c_i is the lower asymptote parameter (also known as the guessing parameter), and
D is a scaling factor (usually equal to 1.702).

The Two-PL model is attained if the guessing parameter c_i is constrained to zero for all items in Equation 1 above.

2. The Generalized Partial Credit Model

Under the GPCM model (Muraki, 1992), the probability, P_{jik}, of the categorical response k on item i for an individual j with ability θ is given by the familiar logistic function:

$$P_{jik}(\theta_j) = \frac{\exp\left[\sum_{v=1}^{k} Z_{iv}(\theta_j)\right]}{\sum_{c=1}^{m} \exp\left[\sum_{v=1}^{c} Z_{iv}(\theta_j)\right]} \tag{2.1}$$

$$Z_{ik}(\theta_j) = Da_i(\theta_j - b_{ik}) = Da_i(\theta_j - b_i + d_k) \tag{2.2}$$

where
a_i is a slope parameter (or item discrimination),
b_i is an item-location parameter (or item difficulty),
b_{ik} is an item-category difficulty parameter, where b_{ik} equals to b_i-d_k and
d_k is a step difficulty.

The partial credit model (PCM, Masters, 1982) is a restricted form of the GPCM model, obtained by further constraining the item discrimination index a_i to be identical for all items. When all items have only two categories to be chosen, the GPCM becomes a Two-PL model.

Only m_i-1 item-category parameters can be identified when the number of response categories is m_i. The item-category difficulty, b_{i1} (or step difficulty, d_1) of the first category on each item is arbitrarily set to zero and the location constraint of

$$\sum_{k=2}^{m_i} d_k = 0 \tag{2.3}$$

is imposed to eliminate indeterminacy (Muraki, 1992).

Figure 1 is GPCM item categories probability curves when a=1, b_{i2}=.5, b_{i3}=1 and b_{i4}=1.5. This plot shows that the item-category parameters are the points on the θ scale at which the item-category plot of $P_{j,k-1}(\theta)$ and $P_{jk}(\theta)$ intersects. For instance, the first and second category curves intersect at 0.5 on the θ-scale, at which the value of .5 represents the second-category parameter (b_{i2}). The values of item-category are interpreted as the relative difficulty of category k in comparing other categories within an item. The item-category parameters equal a constant (the item location parameter for this item) minus the corresponding step parameters that are not necessarily ordered sequentially within an item (Muraki, 1992). This also technically implies that no constraint for order of the item-category (1, 2, 3, … k) parameters is required.

When we expect examinees taking a test to find it easier to reach a lower level k-1 in an item than to reach a higher level k , the frequency of examinees, F_{ik}, reaching a lower level k-1 can be, in general, greater than the frequency reaching a higher level k in an item (refer to Masters, 1982).

$$F_{i1} \geq F_{i2} \ldots \geq F_{ik} \tag{3}$$

Because the frequency, F_{ik}, is a sufficient statistic for estimating b_{ik} under the PCM model, a member of the Rasch model family, the estimates for the item-category parameters must always be ordered.

$$b_{i1} \leq b_{i2} \ldots \leq b_{ik} \tag{4}$$

For the GPCM, ordered item-category parameters, b_{ik}, are preferred, but are not required. The order of the item-category values may produce lower SEs of item estimates as demonstrated by the three-dimensional graphs later.

B. Principles for Calculating the Asymptotic SEs Without Test Data

As indicated previously, a variety of legitimate factors can cause errors in the parameter estimates. This section focuses on those factors that are associated with estimating asymptotic SEs in the ML parameter estimates (refer to Hambleton, Swaminathan & Rogers, 1993; Stocking, 1990; Thissen & Wainer, 1982).

The method of sampling subjects can substantially affect the magnitude of the SE in the estimation of a parameter and can not be ignored due to the "sample free feature" in IRT. As a matter of fact, for estimating item difficulty, easy items and hard items are not well estimated if ability is a bell-shaped distribution centered around the mean ability level. For estimating the guessing parameter, only low-ability groups are informative. When there are few low-ability examinees available to estimate the guessing parameter of a relatively easy item, such a condition makes its SE very large. In addition, the large covariance between the guessing estimate and the difficulty (location) estimate then "causes this uncertainty to move partially to the estimate of location" (Thissen & Wainer, 1982, p 403). For estimating item discrimination, examinees whose abilities are above and below the item difficulty are informative (Stocking, 1990). Such a condition may be found in a broad distribution of abilities relative to the parameters being estimated. In short, the distribution of abilities is associated with the SEs of item parameter estimates.

The magnitude of the value of an item parameter itself may have an effect on its SE value. On the average, the hard items and easy items have larger standard errors; as do the high and low discrimination items. When the distribution of abilities is bell-shaped, the SE of an item difficulty associated with a high discrimination parameter, is lower than the same item difficulty associated with a low discrimination parameter (see Figures 2, 3 and 4, Thissen & Wainer, 1982). Thus, the combination of a set of item parameters for an item should be taken into account when modeling the SE in the estimation of parameter estimates.

The sample size is also a substantive factor affecting the SEs of parameter estimates. The larger the sample size; the lower the SE.

In summary, it should be stated that sample size, the shape of the examinees' ability distribution and the characteristic of test items can each cause differences in the errors in the parameter estimates. A mathematical expression for this relationship is given in Appendix A for dichotomous and polytomous item response data.

III. Methodology

A. Three-D Graphical Presentation to the Analytically Asymptotic SEs

As indicated, the combination of item characteristics (e.g., difficulty, discrimination, and guessing) is one of the key factors to effect the magnitudes of SE for each of various item estimates. The three-D graphical presentation will be used to illust?ate this issue.

The mathematical functions presented in Thissen and Wainer's study (1982) and in Appendix A clearly point out that the AEA-SEs of a set of ML item estimates for an item are a function of the IRT model, the sample size and the shape of the examinees' abilities. Here, the latent trait distribution is assumed normal, N(0,1). In regard to sample size, how large (e.g., 1000, 2000, or 3000) is reasonable for item parameter estimation? The sample size ratio (SSR, sample size to the number of item parameters, refer to De Ayala & Sava-Bolesta, 1999) can be a more objective index to resolve this issue than sample size alone and was employed in this study. For instance, results from De Ayala and Sava-Bolesta (1999) indicated that the SSR of 10:1 can yield reasonably accurate parameter estimates for the nominal response model (Bock, 1970) when examinees' abilities are normally distributed. This standard (SSR=10:1) is used in this study.

When SSR equals 10:1 and a 40-item test is constructed, the sample sizes of 800 and 1,200 are required for the Two-PL and the Three-PL models. Similarly, the sample size 2000 is necessary for the 40 four-category scored GPCM items.

B. Comparisons between the AEA-SEs and the EMB-SEs

1. Test Data Generation

The simulated item parameters were from the Algebra Assessment, designed and scored by the Educational Testing Service (1998, Algebra End-of-course Examination Report). This test consists of 24 multiple-choice items, 8 short-response dichotomous-scored items and 10 constructed response items (3 three-category items, 3 four-category items and 4 five-category items).

The simulated test data were generated by the computer program, RESGEN2.1 (Muraki, 1997). The computer program, PARSCALE (Muraki & Bock, 1996) was used for item calibration. The computer program, EQUMIXED (Li, 1999), was used for computing the item linking coefficients for a mixed-format test. These linking coefficients were then used to convert the scale of the parameter estimates to the scale defined by the true parameters. The sample was set to 1290 to meet the requirement of SSR=10:1. The number of replication was set to 50.

2. Calculating the Empirical MMLE/Bayesian SEs

The EMB-SE was obtained by using the following steps.

(1). Generate a test dataset by the procedures indicated previously;

(2). Simultaneously fit the Two-PL, Three-PL and GPCM models to appropriate item responses and calibrate item parameter estimates, using the MMLE/Bayesian estimation method;

(3). Transform the metric of the estimated parameters to the one defined by the true parameters;

(4). Repeat steps 1 through 3 a large number of times, which results in a large number of estimates for each individual parameter, and

(5). Calculate the BIAS and RMSE (root mean squared error) for each of the parameter estimates by the formulas shown below.

$$\mathrm{BIAS}(H_i) = \frac{\sum_{i=1}^{r}(\hat{H}_i - H_i)}{r} \qquad (5)$$

$$\mathrm{RMSE}(H_i) = \sqrt{\frac{\sum_{i=1}^{r}(\hat{H}_i - H_i)^2}{r}} \quad \text{and} \qquad (6)$$

where H_i is the true item parameter, $\hat{H}_i$ is the corresponding estimated item parameter, and r is the number of replications, in which r equals 50 in this study.

RMSE is a measure of total error of estimation that consist of the systematic error (BIAS) and random error (SE). These three indexes are related to each other as follows:

$$\mathrm{RMSE}(H_i)^2 \cong \mathrm{SE}(H_i)^2 + \mathrm{BIAS}(H_i)^2 \qquad (7)$$

The empirical MMLE/Bayesian SE of an item estimate is approximately estimated by:

$$\mathrm{SE}(H_i) \cong \sqrt{\mathrm{RMSE}(H_i)^2 - \mathrm{BIAS}(H_i)^2} \qquad (8)$$

Another method to estimate the SE of an item estimate is to directly compute the standard deviation of the 50 item estimates obtained from Step 4. This method was not adopted in this study because the SE index as well as BIAS and RMSE indices are all important indices to be used for evaluating the measurement errors of item estimates.

3. Calculating the AEA-SEs

The same 42 sets of item estimates were also used to generate the AEA-SEs of item estimates. The estimated posterior distribution of abilities reported from the PARSCALE output during the item calibration process was used to define the latent distribution of abilities. The same sample size, 1290, used to generate item response data, was used here.

4. Data Analysis

Descriptive statistics of the SE Index of item parameter estimates for AEA and EMB were calculated. Because the same set of item parameters was repeatedly used for estimating their SEs across research conditions, the Log[SE] (Harwell, Stone, Hsu & Kirisci, 1996) of each of various item parameter estimates was treated as a repeated- measure across research conditions. A t-test for dependent observations was then performed to compare the impact of the estimation method on the precision of SE estimates for item parameters.

The Pearson correlation coefficient between AEA-SE and EMB-SE measures, across test items, was calculated for each of various item estimates. Similar calculations were performed for the correlations between BIAS and AEA-SE, across test items, as well as BIAS with EMB-SE. The plots of SEs of item estimates as a function of true item parameters for the AEA and EMB were graphed.

IV. Results

A. Using 3-D Presentation to Explore the AEA-SEs of Item Estimates

1. Three-PL Logistic Model

For the Three-PL model, figures 2a and 2b present plots of SEs of item discrimination and difficulty as the bivariate function of both item estimates while the guessing parameter was set to a constant of 0.25. Figure 2a is for the SEs of bs. It indicates that easy items are more likely to have more measurement error as highlighted by Thissen and Wainer (1982). The SEs of the difficulty parameters can reach an unacceptable magnitude, for instance, larger than 2 for an item with a set of parameters (a =1.3, b = 0 and c =.25).

Figure 2b turns its focus on the SEs of a-parameters. This plot raises an interesting issue that the SE of an a-parameter becomes very high when the same item's b-parameter value is extreme (e.g., too hard or too easy). Fortunately, this combination of item parameters (high discrimination with very high or very low difficulty) does not usually occur in real testing situations because the likelihood of producing a high discrimination parameter seems to be rare for a too hard or too easy item.

Figure 2c presents SEs of the guessing parameters as the bivariate function of the guessing and difficulty estimates, when the a-parameter was set to a constant of 1.5. This plot clearly indicates that the problem of estimating the guessing parameters occurs when an item is relatively easy.

2. Two-PL Logistic Model

For the Two-PL model, figures 3a and 3b present plots of the SEs of item estimates as the bivariate function of both difficulty estimates and discrimination estimates. Figure 3a is for the

SEs of bs. Comparing 3a with 2a (similar plot for the Three-PL model) shows that the problem of ML item difficulty estimates for the Two-PL was not as serious as for the Three-PL. In general, the magnitudes of SEs of bs can never be greater than 0.7 under any circumstance. When the value of SE for an item's b-parameter reaches .7, it suggests that this b-parameter's combination with the a-parameter is one of unrealistic combinations of a set of item estimates (e.g. hard item with high discrimination parameters).

The value of .7 also connotes a special meaning for the Three-PL model. When the Three-PL c-parameter is perfectly estimated, the c-parameter estimate no longer affects the estimates of b-parameter and a-parameter. Under this circumstance, the Three-PL model can be analogous to the Two-PL model so that the maximum possible value of the SE for the b-parameter is expected to be about 0.7, under SSR = 10:1. Therefore, when a SE of an item's b-parameter is larger than .7, from the analytic SE perspective, there are three possible reasons: (a). this item's c-parameter was poorly estimated, (b), a worse or unrealistic combination of a set of item estimates (a and b) for this item occurred, and (c) both factors combined.

Figure 3b is for the SEs of a-parameters. This plot is very similar to the Figure 2b (generated by the Three-PL model), except that the magnitudes of SEs are relatively smaller in the Two-PL model than the Three-PL model for the high-discrimination hard items.

3. Generalized Partial Credit Model

Figure 4a is graphed to explore the relationship between the SEs of the category-parameter (b_{i2}) and the bivariate function of category-parameter (b_{i2}) and the discrimination parameter, when the category-parameters b_{i3} and b_{i4} were set to constants of -1 and 0. On the spot where the values of b_{i2} are less than –1, these category-parameters are ideally ordered ($b_{i2} < b_{i3} < b_{i4}$), the SEs of b_{i2} become relatively small; otherwise, the SEs of b_{i2} could become enormously large when the category-parameters are not ideally ordered, e.g., $b_{i2} > b_{i3}$ and $b_{i2} > b_{i4}$. Comparing Figure 4a with Figure 4b where the category-parameters b_{i3} and b_{i4} were set to constants of 3 and 4, the SEs of most b_{i2} estimates (Figure 4b) become relatively small. The fact that their item-category estimates are ordered might be one of the causes.

With respect to the SEs of the discrimination estimates, Figure 4c was plotted with the same conditions as those used for Figure 4b. This figure shows that the higher values of the GPCM discrimination parameter, the higher SE produced, especially with the condition of extreme b_{i2} values. This phenomenon was also found in the Three-PL and Two-PL models.

On the whole, the Three-D plots of SEs of item estimates for the Three-PL, Two-PL and GPCM models, imply that models without the guessing parameter are more likely to have item parameters that are more precisely estimated.

B. Comparisons Between the AEA-SEs and the EMB-SEs

1. The Three-Parameter Logistic Model

Figure 5 presents plots of the SE as a function of true parameters for the Three-PL item discrimination (a), the item difficulty (b) and the guessing parameter (c), under the AEA and EMB methods. The Three-PL section on Table 1 shows summary descriptive statistics for the SE, computed across 24 items, for each method.

Figure 5 shows that the results from the EMB and AEA methods were similar except for an extreme case for a set of item estimates for an item. Originally, this set of item parameters (a=.258, b=.113 and c=.318) were calibrated from real test data with sample size equal to 6426 (see Table 2). Intuitively, the SE of b-parameter from the BILOG output is relatively large, .581.

The AEA-SEs of b-parameter were 1.275 and 2.807 for N=6426 and for N=1290. As indicated, the reasons why the AEA-SE of an item's b-parameter was larger than .70 under N=1290 or SSR=10:1, where the factors of poor c-parameter estimate and a worst combination of a set of item estimates (parameters b and a) for an item might cause this to happen. As a matter of fact, Figure 2a indicates this combination (a=.258 and b =.113) is one of worst combinations leading to large SEs of b-parameter. Poor c-parameter estimate will be explained below.

When this set of item parameters were repeatedly estimated 50 times under SSR=10:1 (or N=1290), the replication-based SE (.743), BIAS (.782) and RMSE $(\sqrt{(.743)^2 + (.782)^2} = 1.079)$ for this b-parameter were also relatively large. These advanced analyses for this item seem to imply that when variation of b-parameter estimate is relatively large, larger BIAS value for this b-parameter estimate is expected. When we furthermore examined the correlation between the absolute value of BIAS for b-parameter and its SE index, across 24 items, their correlations were very high, e.g., with AEA-SE = .97 and with EMB-SE = .89 (see Table 3). Similar result was found for the c-parameter. Since the item parameter estimates, a, b, and c, are interrelated, these results suggest that when a larger SE of a parameter, especially for the b-parameter, is found, large BIAS for this parameter will be expected. Consequently, the item estimates for this type of item can not be used for testing situations due to unreliable and inaccurate estimates.

The Three-PL section of Table 1 shows that average SEs of item estimates (a, b and c), across the 24 items, from the AEA method were slightly larger than the corresponding indices reported for the EMB method. The dependent t statistic (Table 1) for the Log[SE] of the a-parameter showed statistically significant difference between the AEA and the EMB methods. It seems that there is no practical meaning for this statistically significant difference since this difference (.01) is minimal. No significant differences were found for the parameters b and c.

The correlation coefficient between AEA-SE and EMB-SE measures, across the 24 a-parameters, were .90, .89 and .91 for the parameters, a, b and c.

These results from figure 5 and Table 1 imply that AEA and EMB produce very similar SEs of item estimates, except that AEA might produce some relatively large SEs for some sets of Three-PL item parameters, as demonstrated in Figures 2a, 2b and 2c. If this exception occurs, a close examination for those sets of item parameter estimates is needed because their corresponding BIAS and RMSE values might also be relatively large.

2. The Two-PL Logistic Model

Figure 6 presents plots of the SE as a function of true parameters for the Two-PL item discrimination (a), and the item difficulty (b), under the AEA and EMB methods. The Two-PL section of Table 1 shows summary descriptive statistics for SE, computed across 8 items, for each method.

Figure 6 shows that the results from EMB and AEA methods were similar. The Two-PL section of Table 1 shows that the average SE of parameters a and b from the AEA method was slightly lower than those produced from the EMB. No significant differences (Table 1) were found. The correlation coefficients between AEA-SEs and EMB-SEs, across 8 items, were .97 for both the parameter a and b estimates.

3. The Generalized Partial Credit Model

Figure 7 presents plots of the SE as a function of true parameters for the GPCM item discrimination (a), and the item-category difficulties (b_{ik}), under the AEA and EMB methods. The GPCM section of Table 1 shows summary descriptive statistics, computed across all GPCM items, for the SE for each method.

Figure 7 shows that the results from EMB and AEA methods were similar. Table 1 shows that average SE of parameters a and b from the AEA method was slightly lower than those produced from using EMB. Significant differences were found (Table 1), except for the item – categories, b_{i4} and b_{i5}. The correlation coefficients between AEA-SEs and EMB-SEs, across test items, were .97, .93, .94, .99 and .99 for a-parameter, and item-category parameters, b_{i2}, b_{i3}, b_{i4} and b_{i5}, respectively.

Two-PL is a special case of the GPCM model. In general, AEA consistently produced lower SE values for each of the various item estimates than EMB for these two models. Table 1 shows no significant differences in SEs for the Two-PL model were found, but several significant differences occurred for the GPCM model.

As indicated in Equations 3 and 4, when a GPCM item has the characteristic: easier success at a lower level than at higher level for examinees, the item-category difficulties (or frequencies of examinees reaching any category within an item), generally, will be ordered. The AEA Three-D graphs presented previously imply that an item with this characteristic could be one of the causes of stabilized item-category estimates. The values of EMB-SEs of item-category parameters from some items also support this point. For instance, the set of item-category parameters for a GPCM item used in this study were b_{i2}=2.83, b_{i3}= -0.69, b_{i4}=6.40 and b_{i5}=-2.32. These item-category parameters were not ordered. Their corresponding EMB-SEs (or AEA-SEs) were relatively larger, 0.26 (0.22), 0.20 (0.18), 0.64 (0.64) and 0.62 (0.63) (note: values in parentheses were AEA-SEs).

In another example in which a set of item-category parameters were ordered (b_{i2}=-0.41, b_{i3}= 1.76, b_{i4}=2.41 and b_{i5}=2.91), their corresponding EMB-SEs (or AEA-SEs) were relatively smaller, 0.07 (0.05), 0.11 (0.10), 0.21 (0.17) and 0.39 (0.33).

Considering the Two-PL as a member of the GPCM model, the AEA and EMB, in general, produce very similar SEs of item estimates for the test data evaluated in this study.

V. Applications of Conclusions

A. General Applications

For the practical application, Thissen and Wainer (1982) illustrated how to use the AEA-SE for the determination of the sample size required to yield desired accuracy for any set of item parameters. As demonstrated in Figures, 2, 3 and 4, this type of application should be more meaningful and practical when any unreasonable combination of a set of item estimates for an item (e.g., hard item with high or low discrimination parameter) is excluded. As a matter of fact, this unrealistic combination of item estimates is rarely found in real data. If it occurs, the level of precision for these item estimates would be questionable.

Tabulating the AEA-SEs of item estimates under some conditions (e.g., different sample sizes, levels of item difficulty or discrimination, etc.) is another means to provide test practitioners a sense of the accuracy of parameter estimates, on which SEs of item estimates are yielded under a specific situation (refer to Thissen & Wainer , 1982).

The three-D graphical presentation for AEA-SEs of item estimates can be used for detecting the possible problems of the ML estimation method to the IRT models of interest. Using this graphical procedure, Thissen and Wainer (1982), along with this study, have pointed out that the widely-used Three-PL has potential problem obtaining accurate ML item estimates. Fortunately, this is not the case for the current popular estimation method, MMLE/Bayesian, that minimize this problem for the Three-PL model.

The three-D graphical analyses for the GPCM suggests that when constructing the GPCM test items, easier success at a lower level in an item than at a higher level for examinees needs to be considered . This principle, in general, makes the item-category difficulties ordered and that may decrease the variation of item-category estimates.

AEA tends to produce relatively larger SEs for some combinations of Three-PL item parameters (e.g. a=.258, b=.113 and c=.318; also refer to figure 2) than the replication-based approach or BILOG. Although this fact is AEA's drawback, we might make use of this fact to help us identify questionable Three-PL item parameter estimates when we have trouble deciding whether the numerical SE values generated from the BILOG-output are large enough to be identified as unstable and inaccurate item estimates. More specifically, when we find that some sets of item estimates, along with large SEs, generated from an IRT computer software, we might also calculate the corresponding AEA-SEs for these sets of item estimates. If the magnitudes of AEA-SEs of b-parameters are large (e.g., larger than .7 under SSR=10:1), we might suspect that the set of item estimates are unstable and inaccurate and can not be used in real testing situations.

For the test data examined in this study, we also find that Three-PL item estimates with larger AEA-SEs also have larger BIAS values. This finding suggests that the AEA might be a nice tool to identify which sets of item estimates are contaminated with large BIAS error. The issue of how well the AEA approach can be used to flag biased item estimates for the Three-PL model needs to be closely examined.

B. Application to Item-Linking Studies

Several unidimentional IRT equating methods for placing test items, separately calibrated from different test forms, on the same metric exist under a common-item linking design (Vale, 1986) in which tests containing a set of common items are administered to two groups of examinees. They can be grouped as, the mean/sigma method (Marco, 1977), the mean/mean method (Loyd & Hoover, 1980), the item characteristic curve method (Haebara, 1980), the test characteristic curve method (Stocking & Lord, 1983), the minimum chi-square method, Divgi, 1985), and the numerical integration method (Zeng & Kolen, 1994). If we attempt to explore which method is the most robust to the random error of item estimates for the polytotomus-scored test. This type of research is very time-consuming when random error of item estimates is manipulated by the replication approach described in the previous section on Methodology. In contrast, if the random error of item estimates is manipulated by the analytic-based approach, researchers can take advantage of its significant features described previously.

Conceptually, an item estimate has three components, true item parameter value, a random error and bias. When bias is assumed to be zero and a set of true item parameters for an item is given, the procedures of adding "reasonable numerical values as random errors" to this set of true parameters to form a set of item estimates are illustrated below.

When the latent trait distribution of 1000 examinees' abilities is distributed as N(0,1), the variance-covariance matrix, **V** shown below, for a set of item parameters, a=1.2, b=0.5 and c=0.2, can be predicted using Equation 13 :

$$\begin{array}{cc} & \begin{array}{ccc} a & b & c \end{array} \\ V = & \begin{bmatrix} .0239 & .0058 & .0034 \\ .0058 & .0067 & .0021 \\ .0034 & .0021 & .0012 \end{bmatrix} \end{array}$$

The square roots of the diagonal elements of the matrix **V** are the asymptotic standard errors of the parameters. They are .155, .082 and .034 for the parameters, a, b and c.

When a matrix, **E** shown below, is randomly generated from MVN(**0**, **V**) using the computer software, MATLAB (The MathWorks, Inc, 1999), random errors for the parameter estimates, a, b and c, are the diagonal elements of matrix **E**.

$$\begin{array}{cc} & \begin{array}{ccc} a & b & c \end{array} \\ E = & \begin{bmatrix} .-.0901 & .0051 & -.0134 \\ .0051. & -.0302 & .0222 \\ -.0134 & .0222 & .0493 \end{bmatrix} \end{array}$$

The simulated item estimates for the set of true parameters a=1.2 , b=0.5 and c=0.2 are: a= 1.2+ (-0.0901), b=0.5+(-0.0302)., and c=0.2+0.0493.

It is noted that matrix E is randomly generated from the MVN(0, **V**) so that the values of its elements vary across replications. Therefore, the simulated item estimates for the set of true parameters a=1.2 , b=0.5 and c=0.2 will be changed, along with the changes of the error matrix **E**. Theoretically, when a large number of replications is conducted, the standard deviations of the simulated item estimates, a, b and c, will be close to the expected SEs of parameters a, b and c. They are .155, .082 and .034. This type of modeling measurement errors of item estimates is much easier to employ for some item-linking studies. One research example conducted by Li and Lissitz (2000) is that when several multidimensional IRT item-linking methods exists, we attempt to examine which MIRT item-linking method is relatively less sensitive to the random (or sampling) errors of item parameter estimates. The above procedures of modeling random errors can be incorporated in the following procedures for this type of study:

1. Choose a set of true item parameters for the base test .

2. Assume item linking coefficients are known and generate a set of true item parameters for the linked test by using these linking coefficients.

3.1. Model random errors for the base-test item parameters. Each simulated item estimate from a set of parameters of an item is computed by summing the expected random error and the corresponding true item parameter. Expected random error was generated as a random value using the method outlined above.

3.2. Model measurement errors for the linked-test item parameters, using the method outlined in Step 3.1. It is noted that Step 3 is an alternative method to predict the random errors of item estimates. Comparing the AEA approach with the replication approach to modeling measurement errors of item estimates, the AEA approach will save an enormous amount of time and energy in test data generation and item calibration for some types of research topics.

4. Estimate the equating coefficients based on two sets (base and linked) of item parameter estimates.

(5) Repeat Steps 3.1. 3.2 and 4 a large number of times, which results in a large number of estimates for each individual item-linking coefficient; and Calculate the BIAS (average

difference between estimated and true values and RMSE (root mean squared error) of the item-linking coefficient estimate.

Another research example conducted by Li, Lissitz and Yang (1999) examined whether the principle of matching test characteristic surfaces between the base and linked tests is appropriate for mixed format tests for finding the linking coefficients. Since the AEA had been employed for modeling errors of item estimates, an enormous number of replications (1000) for each research conditions became available. Therefore, the sampling distribution for each equating coefficient might be more accurately estimated as the number of replications increases to 1000. However, a unexpected result indicated in the introduction can occur. Table 4, excerpted from that study, presents the average value of the Three-PL item parameters for the set of 10, 15 or 20 test items and the corresponding average AEA-SE. The fact that the longer test (15 or 20 items) had larger average AEA-SE than the shorter (10 items) test caused this unexpected result to occur. Consequently, when using the AEA approach to model random errors of item estimates, interpreting the research results should be done cautiously.

Considering the MMLE/Bayesian as a standard method for item estimates, the analytic approach to compute the SEs of ML item parameter estimates, in general, underestimates the variation of item estimates for the GPCM model. For the Three-PL model, the analytic approach tends to overestimate variation of item estimates. Although those differences can be minor in most combinations of sets of item estimates, some extreme differences can occur, especially for the Three-PL model. Based on results from this study, the extreme cases most often came form those with unreasonable combinations of sets of item estimates. Those cases will produce larger measurement errors and should be excluded from the studies.

Using AEA for modeling random errors of item estimates has its theoretical limitations. As indicated, measurement errors of item estimates are assumed to be distributed as **MVN** ($\underline{S}$, **V**). AEA is the analytic approach to model the "units of measurement errors for item estimates" (known as SEs of item estimates, associated with the matrix **V**). Besides that, modeling the "points of origin of measurement errors for item estimates" (known as the BIAS of item estimates, indicated by the vector $\underline{S}$) is another key issue to be considered. Although we might assume $\underline{S}$ to be $\underline{0}$, for simplicity, ML is a biased estimator (Anderson & Richardson, 1979) and the degree of bias depends upon the sample size. This issue of modeling $\underline{S}$ needs to be further explored in the future for better prediction of measurement errors of item estimates. Up to this point, when the bias of item estimates may have a strong effect on the research topics being investigated, the AEA approach to modeling measurement error is not appropriate.

C. Application on Estimating Standard Errors for IRT-true Score Equating Scores

As indicated, without using examinees' responses on the test, the expected SEs of item estimates can be predicted. Similarly, the expected SEs of IRT-true score equating scores (Lord, 1982) can also be predicted without using examinees' response patterns on the test. In other words, if a new test is generated from an item pool and edited with some anchor items used in the previous (old) test form, a look-up table for converting the new test scores into the corresponding old-test scores can be generated. After that, the SE for each of the new test scores can be predicted when the information of the AEA-based var-cov matrix of item estimates (see Equation 13) is incorporated with Lord's formula (1982). It is noted that Lord used the observed var-cov matrix of item estimates, generated from a real test data, rather than the analytic one. The detailed procedures on how to compute the SEs of IRT-true score equating can be found in Lord's study (1982).

Figure 8 is used to demonstrate how similar the formula-based SEs of test scores are with those produced by the 500-replication bootstrap method under IRT chained true score equating, Tsai, 1998). The bootstrap approach (Kolen & Brennan, 1995) was used to compute SEs of IRT equating for the nonequivalent-group and common-item design. The possible score range was 0 through 150 for the test used in his study. Since approximately 95% of the examinees had raw scores in the range of 67 through 125, Figure 8 only presents the similarity of SEs of scores calculated by both approaches for this score range. Since this is a licensure test, it is important to check the precision of equating near the passing score, 88. The SEs at this passing score are approximately .51 and .49 for the analytic and the bootstrap methods. Part of the reason why the analytic method consistently produced larger SE than the bootstrap did is that the analytic var-cov matrix for the Three-PL item estimates, in general, is larger than the corresponding observed var-cov matrix. A new research on the issue of using the AEA-based var-cov matrix for predicting the SEs of IRT-true equating scores is still on-going by the authors.

Finally, the SEs yielded from the AEA and the EMB approaches were very similar in most cases, especially for the GPCM model or models without the lower asymptote parameter. This finding indeed encourages test practitioners and researchers to apply the asymptotic SEs (or var-cov matrix) of item estimates to practical testing situations and in the context of research studies.

References

Anderson, J. A. & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. Technometrics, 21, 71-78.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Baker, F. B. (1992). Item Response Theory: Parameter Estimation Techniques. New York: Marcel Dekker, Inc.

Beaton, A. E. & Zwick, R. (1992). Overview of the national assessment of educational progress. Journal of Educational Statistics, 17, 95-109.

De Ayala, R. J. & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. Applied Psychological Measurement, 23, 3-19.

Gruijter D. N. M. de (1984). A comment on 'some standard errors in item response theory. Psychometrika, 49, 269-272.

Haebara, T. (1908). Equating logistic ability scales by weighted least squares method. Japanese Psychological Research, 22, 144-149.

Hambleton, R. K. & Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park: CA. SAGE Publications, Inc.

Harwell, M. R., Stone, C. A., Hsu, T. , & Kirisci, L. (1996). Monte carlo studies in item response theory. Applied Psychological Measurement, 20, 101-125.

Li, Y. H. (1999). EQUMIXED: A computer program to estimate equating coefficients for mixed-format tests. [Computer software]. Upper Marlboro: Author.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999, April). Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Li, Y. H. & Lissitz, R. W. (2000). An evaluation of multidimensional IRT linking. Applied Psychological Measurement, 24, 115-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1982). Standard error of an equating by item response theory. Applied Psychological Measurement, 6, 463-472.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rash model. Journal of Educational Measurement, 17, 179-193.

Maco, G. L. (1977). Item characteristic curve solutions to three intractable testing programs. Journal of Educational Measurement, 14, 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika 47, 149-174.

Mislevy, R. J. & Bock, R. D. (1990). BILOG-3: Item analysis and test scoring with binary logistic models. [Computer program]. Mooresvilk: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1997). RESGEN2.1: Item response generator. [Computer program]. Educational Testing Service.

Muraki, E. & Bock, R. D. (1996). PARSCALE (Version 3.): IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. [Computer program] Mooresvilk: Scientific Software.

Nakamura, S. (1996). Numerical analysis and graphic visualization with MATLAB. Upper Saddle River, NJ: Prentice-Hall, Inc.

Tam, H. P. & Li, Y. H. (1997, March). Is the use of the difference likelihood ratio chi-square statistic for comparing nested IRT models justifiable? Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

The MathWorks, Inc. (1999). MATLAB (Version 5.3): The language of technical computing [Computer program]. Natick MA: The MathWorks, Inc.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R.J. Mislevy & I. Bejar (Eds.) Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tsai, T. (1998). A comparison of bootstrap standard errors of IRT equating methods for the common item nonequivalent groups design. (Doctoral dissertation, University of Iowa, 1998). Dissertation Abstracts International, UMI Number 9917609).

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological_ Measurement, 14, 1990.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. Psychometrika, 3, 461-475.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Zeng, L. & Kolen, M. J. (1994, April). IRT scale transformations using numerical integration. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Appendix A

1. For the Dichotomous IRT Models

For an item i, the likelihood of the observed dichotomous responses for N independent examinees is (refer to Thissen & Wainer, 1982):

$$L_D = \prod_{j=1}^{N} P_j^u (1-P_j)^{1-u} \tag{9}$$

where P can be calculated from a three-parameter model, u=1 for correct response; u=0 for incorrect response. The loglikelihood of Equation 9 is

$$\log L_D = \sum_{j=1}^{N} \left[u \log(P_j) + (1-u)\log(1-P_j)\right] \tag{10}$$

The maximum likelihood estimates of each parameter (e.g., a_i, b_i, c_I for the Three-PL model) are located where the partial derivatives of Equation 6 are zero. For ease of expression, ξ represents the three-parameter item parameters (a_i, d_i, c_i). Given a density of θ (e.g. normal distribution, N(0, 1)), for any pair of parameters ξ_s and ξ_t, the negative expected value of the second derivative of the loglikelihood function, Equation, 10, has the form (refer to Thissen, Wainer, 1982),

$$-E\left(\frac{\partial^2 \log L}{\partial\xi_s \partial\xi_t}\right) = N \int_{-\infty}^{\infty} \left[\left(\frac{1}{PQ}\right)\left(\frac{\partial P(\theta)}{\partial\xi_s}\frac{\partial P(\theta)}{\partial\xi_t}\right)\right]\Phi_j(\theta)d\theta \tag{11}$$

where E is the expectation and Q=1-P. Equation 7 requires the derivatives of $P(\theta)$ with respect to its parameters. The numerical approximation of the integral in Equation 11 can be calculated by the Gauss-Hermite quadrature and is presented in Equation 12,

$$I_{3PL}(\xi_s, \xi_t) = N\sum_{q=1}^{q} \left\{\left(\frac{1}{PQ}\right)\left(\frac{\partial P(X)}{\partial\xi_s}\frac{\partial P(X)}{\partial\xi_t}\right)\right\}A(X_q) \tag{12}$$

where X is a quadrature point in the ability dimension, q is the number of quadrature in the ability dimension and A(X) is the corresponding weight of the quadrature. The number of quadrature points for numerical integration are set to 40 in this study.

The partial derivatives of P(X) with its parameters can be resolved using difference approximation (Nakamura, 1996) and substituted in Equation 12 to give a 3 x 3 (for the three-parameter model) information matrix corresponding to the triplet item parameters (a, b, and c). The inverse of that information matrix is the asymptotic variance-covariance matrix of the three parameters and is given in Equation 13. The square roots of the diagonal elements of the variance-covariance matrix are the asymptotic standard errors of the parameters.

$$VarCov_{3PL} = \begin{bmatrix} I_{3PL}(a,a) & I_{3PL}(a,b) & I_{3PL}(a,c) \\ I_{3PL}(b,a) & I_{3PL}(b,b) & I_{3PL}(b,c) \\ I_{3PL}(c,a) & I_{3PL}(c,b) & I_{3PL}(c,c) \end{bmatrix}^{-1} \tag{13}$$

2. For the Polytomous IRT Models

For an item i, the likelihood of the observed polytomous responses for N independent examinees is:

$$L_P = \prod_{j=1}^{N} \prod_{k=1}^{m} P_{jk}^{uk} (1 - P_{jk})^{1-uk} \qquad (14)$$

where P_k can be calculated from a GPCM model, u=1 for the categorical response k; u=0 for responses other than category k. The loglikelihood of Equation 14 is

$$\log L_P = \sum_{j=1}^{N} \sum_{k=1}^{m} \left[u_k \log(P_{jk}) + (1 - u_k) \log(1 - P_{jk})\right] \qquad (15)$$

Similar principles used in the three-parameter model can be applied for the GPCM model to derive the information function when any pair of GPCM item parameter estimates is given. That is:

$$-E\left(\frac{\partial^2 \text{LogP}}{\partial\xi_s \partial\xi_t}\right) = N \sum_{q=1}^{q} \left\{ \sum_{k=1}^{m} \left(\frac{1}{P_k Q_k}\right) \left(\frac{\partial P_k(X)}{\partial\xi_s} \frac{\partial P_k(X)}{\partial\xi_t}\right) \right\} A(X_q)$$
$$= I_{GPCM}(\xi_s, \xi_t) \qquad (16)$$

The inverse of that information matrix is the asymptotic variance-covariance matrix of the four parameters (a_i, b_{i2}, b_{i3},b_{i4}) and is given in Equation 17 for the case of a four-category GPCM model. The square roots of the diagonal elements of the variance-covariance matrix are the asymptotic standard errors of the parameters.

$$\text{VarCov}_{GPCM} = \begin{bmatrix} I_{GPCM}(a,a) & I_{GPCM}(a,b_2) & I_{GPCM}(a,b_3) & I_{GPCM}(a,b_4) \\ I_{GPCM}(b_2,a) & I_{GPCM}(b_2,b_2) & I_{GPCM}(b_2,b_3) & I_{GPCM}(b_2,b_4) \\ I_{GPCM}(b_3,a) & I_{GPCM}(b_3,b_2) & I_{GPCM}(b_3,b_3) & I_{GPCM}(b_3,b_4) \\ I_{GPCM}(b_4,a) & I_{GPCM}(b_4,b_2) & I_{GPCM}(b_4,b_3) & I_{GPCM}(b_4,b_4) \end{bmatrix}^{-1} \qquad (17)$$

Table 1

Descriptive Statistics of SE Index of Item Parameter Estimates, dependent t Tests, and Pearson Correlation Coefficients, for the AEA and EMB methods (N=1290, Replications for EMB = 50)

Method and Parameter	Mean	Number of Items	AEA Mean	AEA Min	AEA Max	EMB Mean	EMB Min	EMB Max	t	r
Three-PL										
a	1.06	24	.20	.10	.49	.19	.08	.39	2.52*	.90
b	1.20	24	.24	.07	2.81	.16	.07	.74	0.20	.89
c	.23	24	.05	.01	.40	.04	.02	.10	0.70	.91
Two-PL										
a	.98	8	.08	.05	.12	.09	.04	.16	-.87	.97
b	1.59	8	.09	.07	.14	.10	.05	.18	-.55	.97
GPCM										
a	.69	10	.04	.02	.11	.06	.03	.13	-8.42***	.98
b_{i2}	1.51	10	.17	.05	.33	.22	.07	.60	-4.07***	.93
b_{i3}	.48	10	.16	.07	.31	.20	.06	.48	-3.89**	.96
b_{i4}	1.94	7	.23	.09	.64	.24	.09	.64	-1.23	.99
b_{i5}	.07	4	.31	.14	.63	.32	.12	.62	0.26	.99

* P < .05; ** P< .01; *** P< .001

Table 2 Measurement Error Components for a set of item parameters

	Parameter	BILOG-SE N=6426	AEA-SE N=6426	AEA-SE N=1290	EMB-SE N=1290	EMB-BIAS N=1290	EMB-RMSE N=1290
a	.258	.039	.078	.172	.170	.190	.255
b	.113	.581	1.275	2.807	.743	.782	1.079
c	.318	.082	.183	.402	.101	.114	.152

Table 3

The Pearson Coefficients between the Absolute value of BIAS Index of Item Estimate and the AEA-SE, as well as EMB-SE Indices (N=1290, Replications for EMB = 50)

	Three-PL			Two-PL		GPCM				
	a	b	c	a	b	a	b_{i2}	b_{i3}	b_{i4}	b_{i5}
AEA-SE	.57	.97	.88	.38	.71	.01	.85	.60	.57	-.09
EMB-SE	.41	.89	.84	.19	.75	.11	.93	.78	.49	.06

Table 4
The Average Value of Item Parameters for the Set of 10, 15 or 20 Three-PL Items and the Corresponding Average SE Value (N=2000)

Item Length	Mean a	Average SE	Mean b	Average SE	Mean c	Average SE
10	.8819	.0845	-.3445	.1347	.1172	.0641
15	.8606	.0858	-.2025	.1517	.1278	.0645
20	.8902	.0866	-.3272	.1489	.1309	.0674

Figure Headings

Figure 1. GPCM item categories probability curves

Figure 2a. The standard errors of item difficulties shown as the bivariate function of both item difficulty and discrimination estimates for the Three-PL model when the guessing parameter is 0.25.

Figure 2b. The standard errors of item discriminations shown as the bivariate function of both item difficulty and discrimination estimates for the Three-PL model when the guessing parameter is 0.25.

Figure 2c. The standard errors of guessing parameters shown as the bivariate function of both item difficulty and guessing estimates for the Three-PL model when the discrimination parameter is 1.5.

Figure 3a. The standard errors of item difficulties shown as the bivariate function of both item difficulty and discrimination estimates for the Two-PL model.

Figure 3b. The standard errors of item discriminations shown as the bivariate function of both item difficulty and discrimination estimates for the Two-PL model.

Figure 4a. The standard errors of item-category difficulties (b_{i2}) shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are -1 and 0.

Figure 4b. The standard errors of item-category difficulties (b_{i2}) shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are 3 and 4.

Figure 4c. The standard errors of item discriminations shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are 3 and 4.

Figure 5a. SE of a as a function of the true a-parameter for the Three-PL model.

Figure 5b. SE of b as a function of the true b-parameter for the Three-PL model.

Figure 5c. SE of c as a function of the true c-parameter for the Three-PL model.

Figure 6a. SE of a as a function of the true a-parameter for the Two-PL model.

Figure 6b. SE of b as a function of the true b-parameter for the Two-PL model.

Figure 7a. SE of a as a function of the true a-parameter for the GPCM model.

Figure 7b. SE of b_{i2} as a function of the true item-category parameter, b_{i2}, for the GPCM model.

Figure 7c. SE of b_{i3} as a function of the true item-category parameter, b_{i3}, for the GPCM model.

Figure 7d. SE of b_{i4} as a function of the true item-category parameter, b_{i4}, for the GPCM model.

Figure 7e. SE of b_{i5} as a function of the true item-category parameter, b_{i5}, for the GPCM model.

Figure 8. Standard Error as a Function of Raw Score for the Analytic and Bootstrap Method

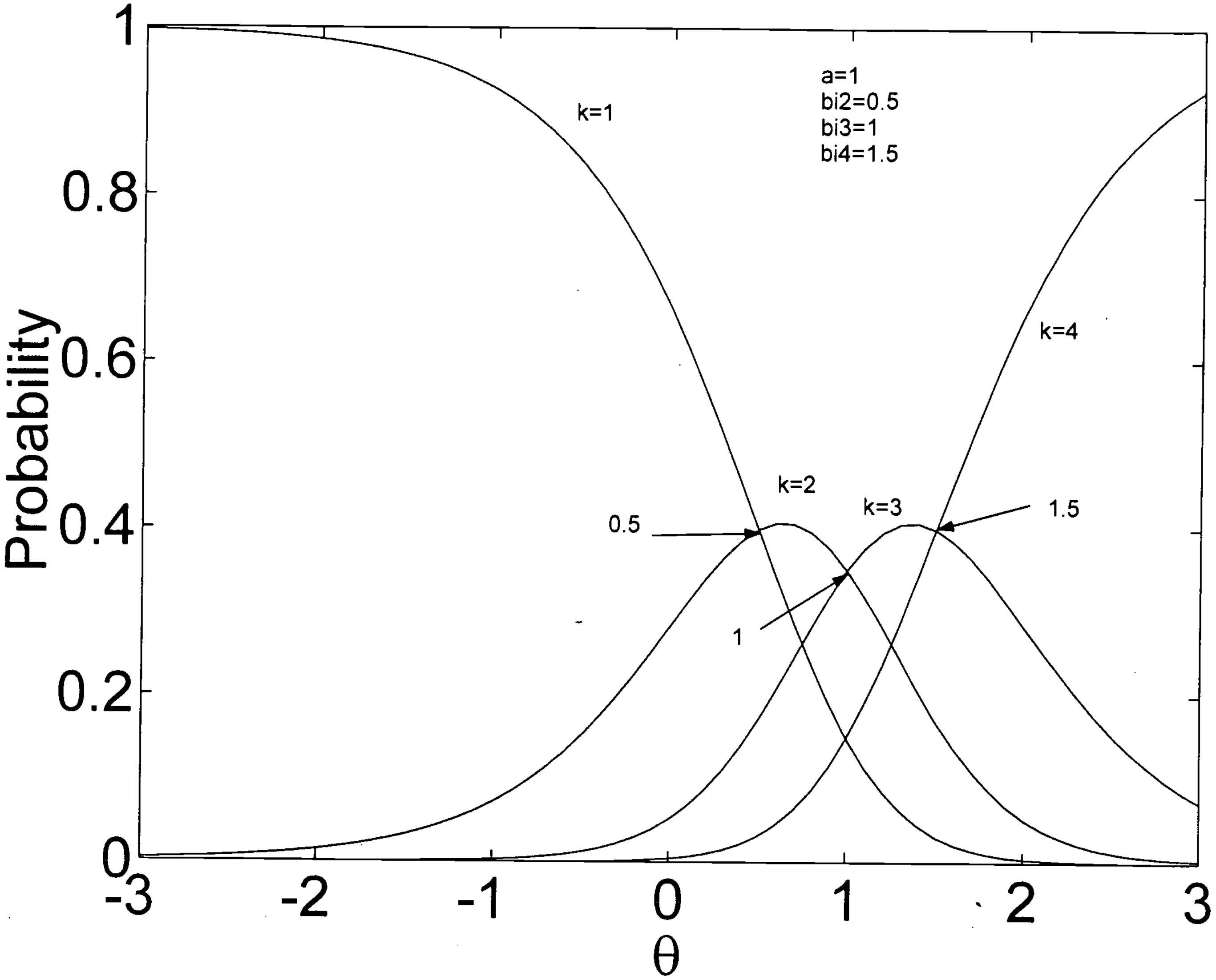


Figure 1. GPCM item categories probability curves

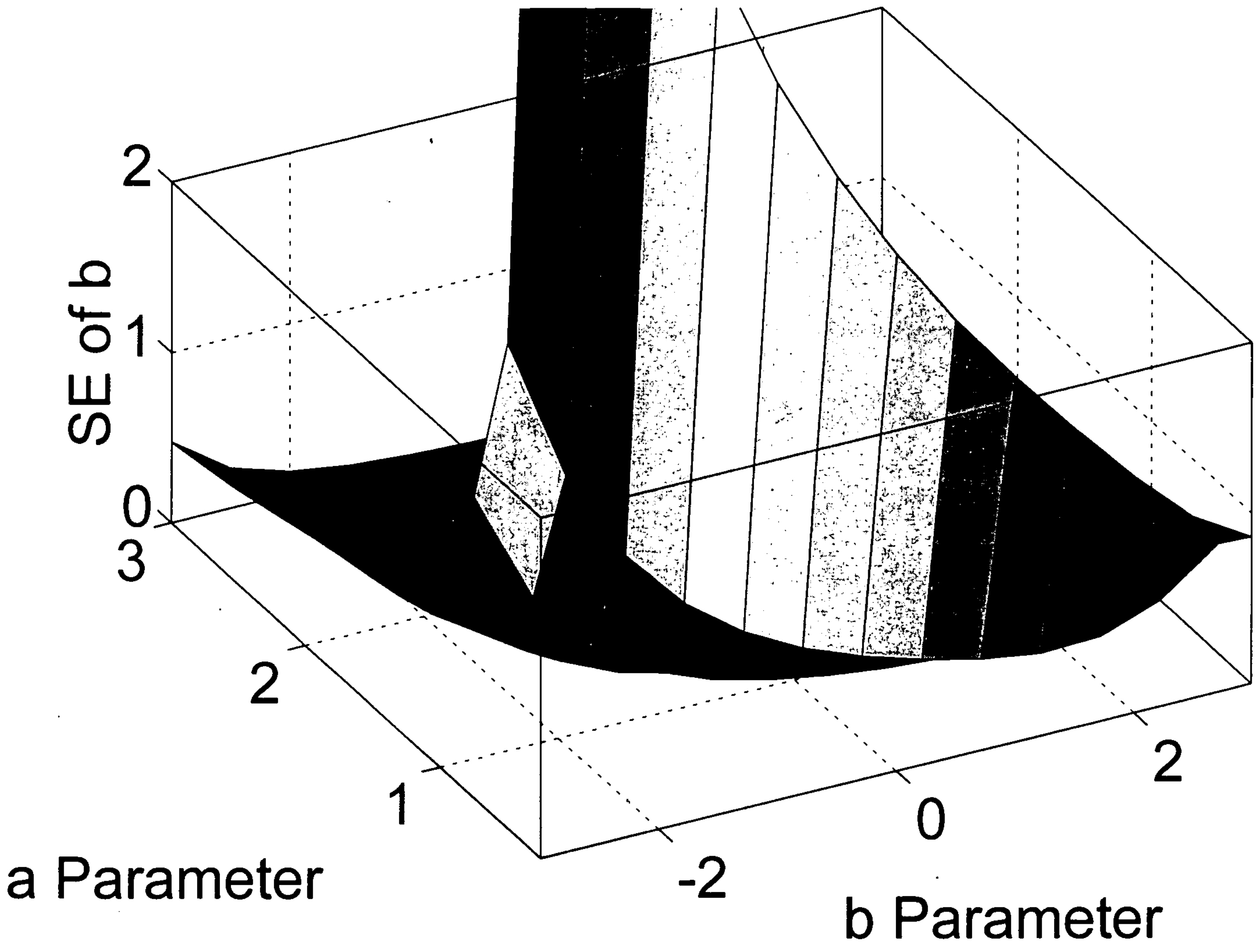


Figure 2a. The standard errors of item difficulties shown as the bivariate function of both item difficulty and discrimination estimates for the Three-PL model when the guessing parameter is 0.25.

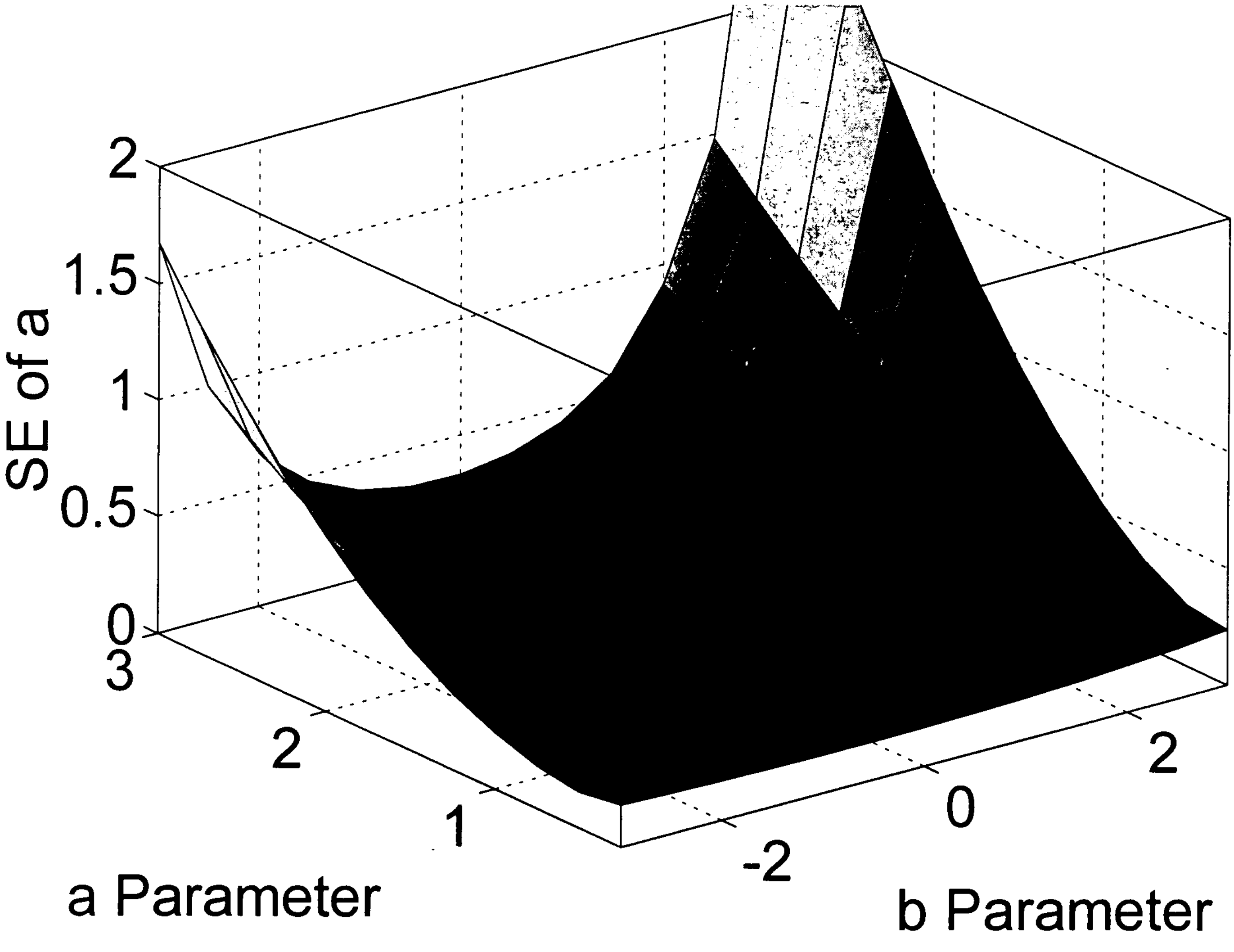


Figure 2b. The standard errors of item discriminations shown as the bivariate function of both item difficulty and discrimination estimates for the Three-PL model when the guessing parameter is 0.25.

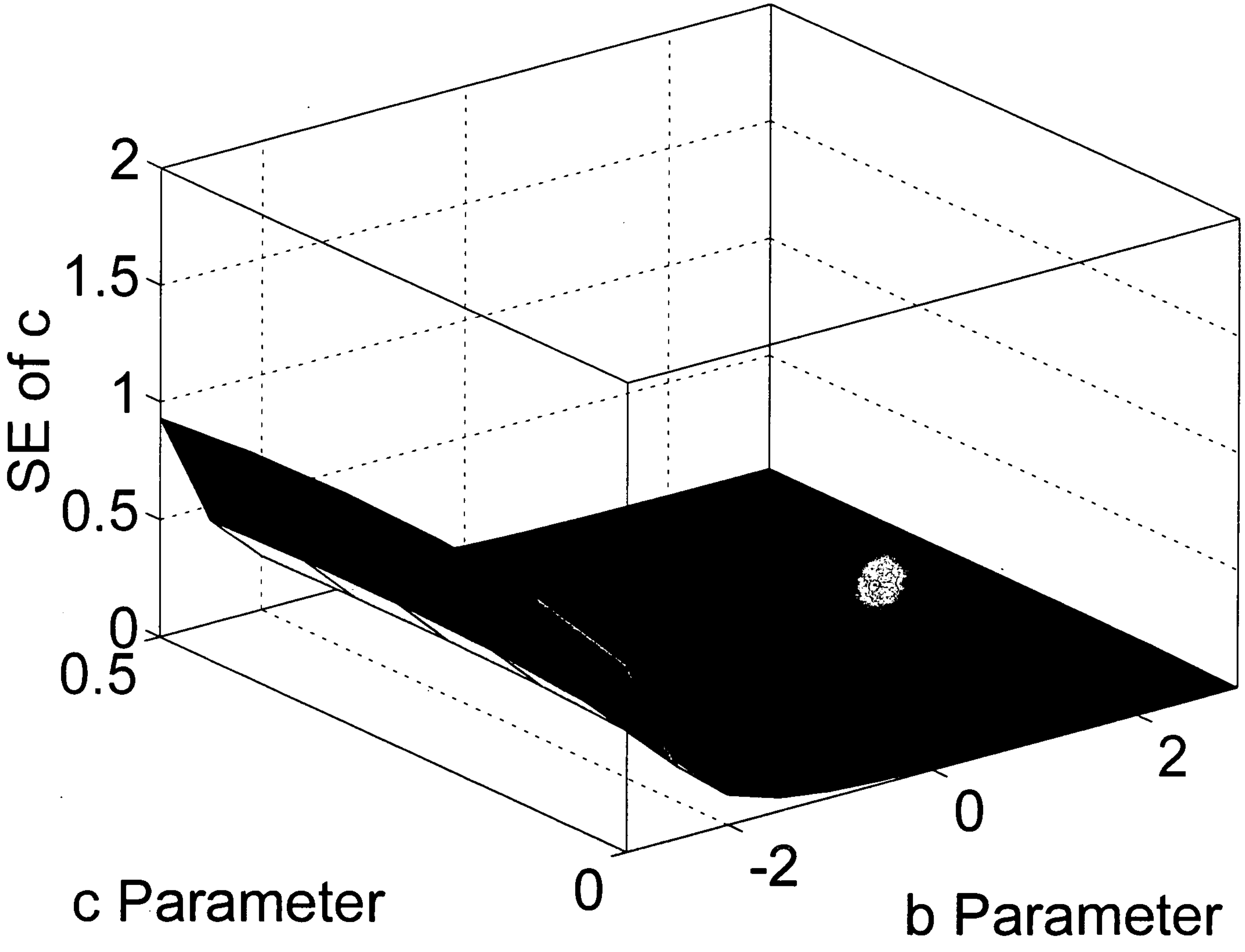


Figure 2c. The standard errors of guessing parameters shown as the bivariate function of both item difficulty and guessing estimates for the Three-PL model when the discrimination parameter is 1.5.

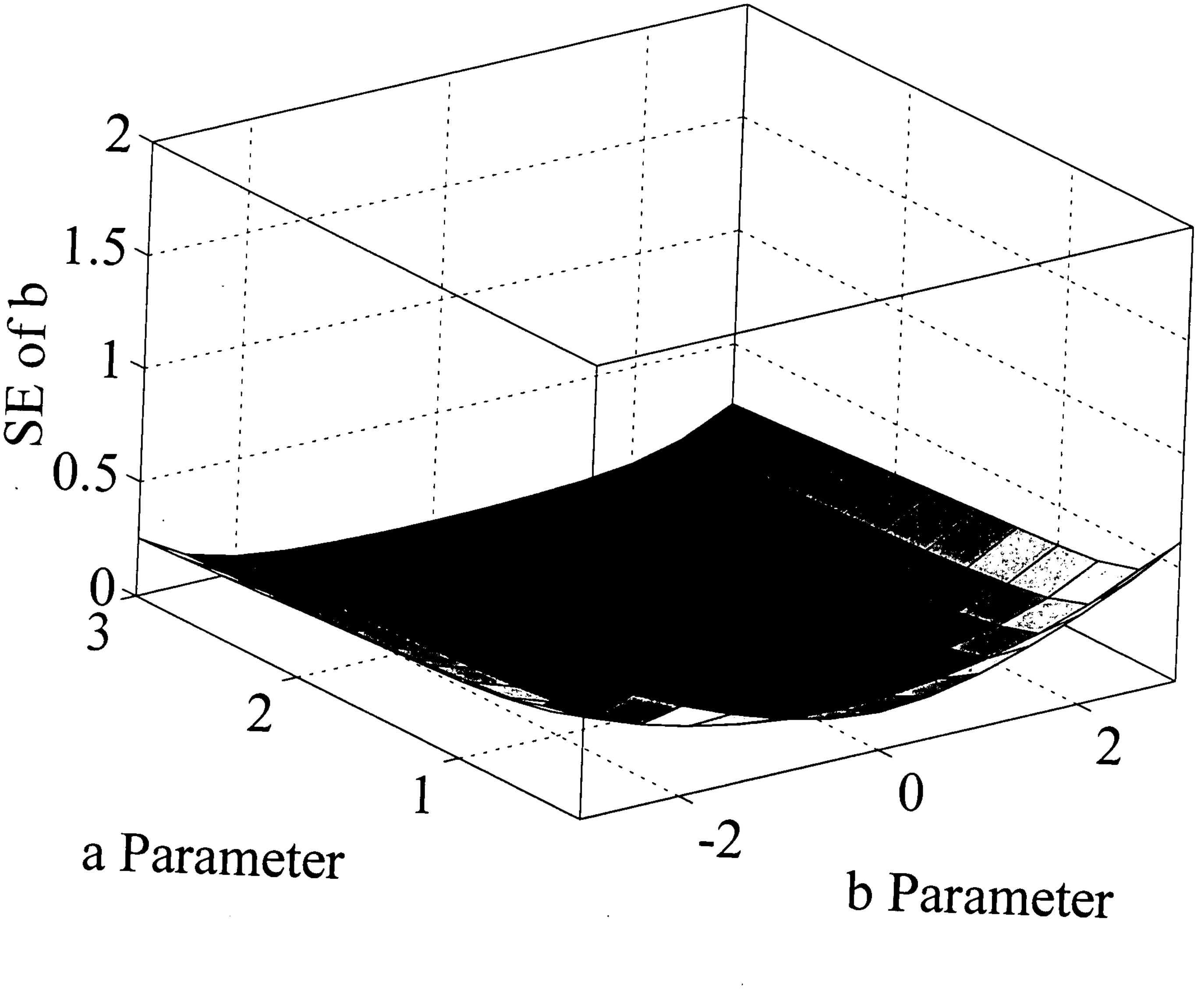


Figure 3a. The standard errors of item difficulties shown as the bivariate function of both item difficulty and discrimination estimates for the Two-PL model.

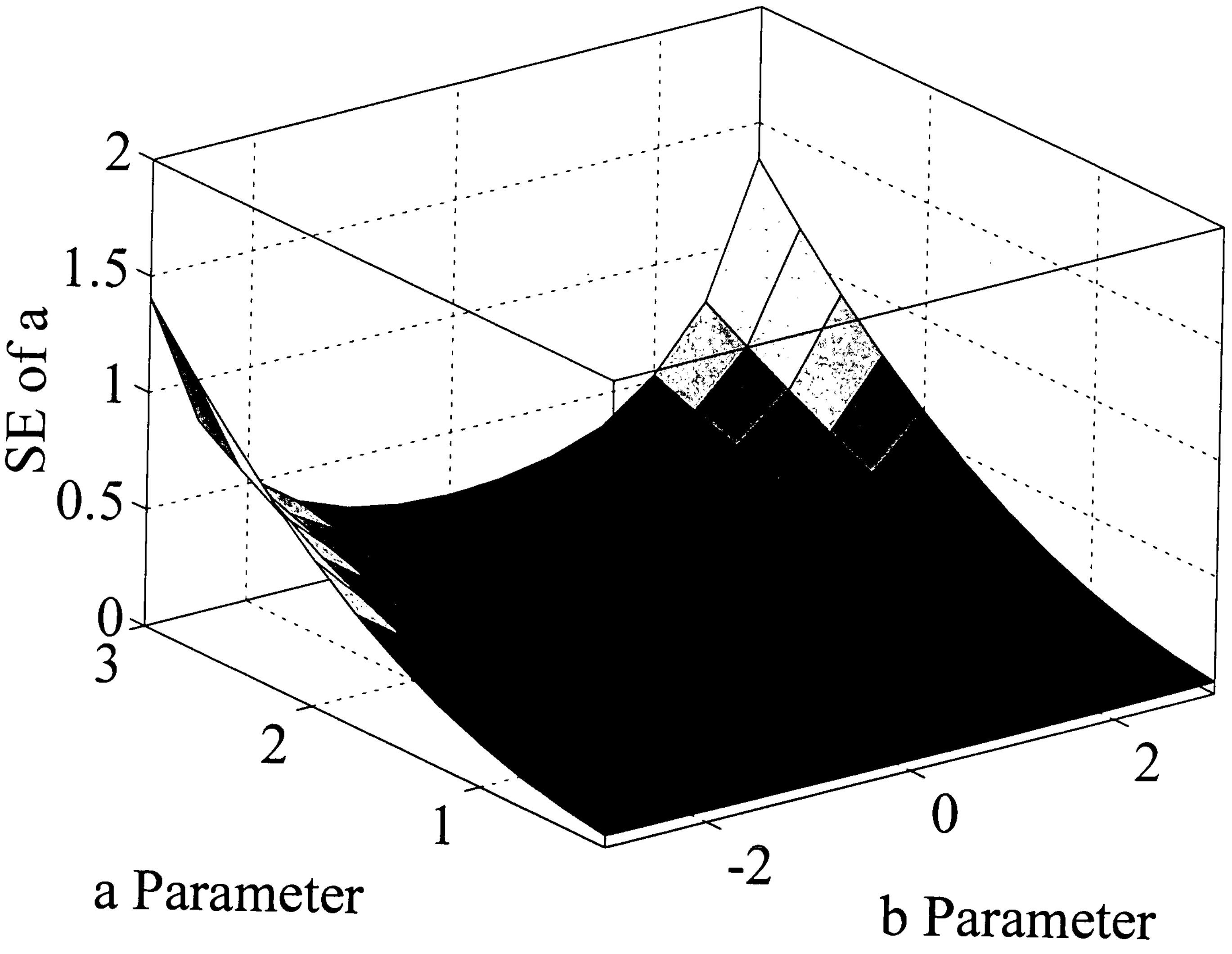


Figure 3b. The standard errors of item discriminations shown as the bivariate function of both item difficulty and discrimination estimates for the Two-PL model.

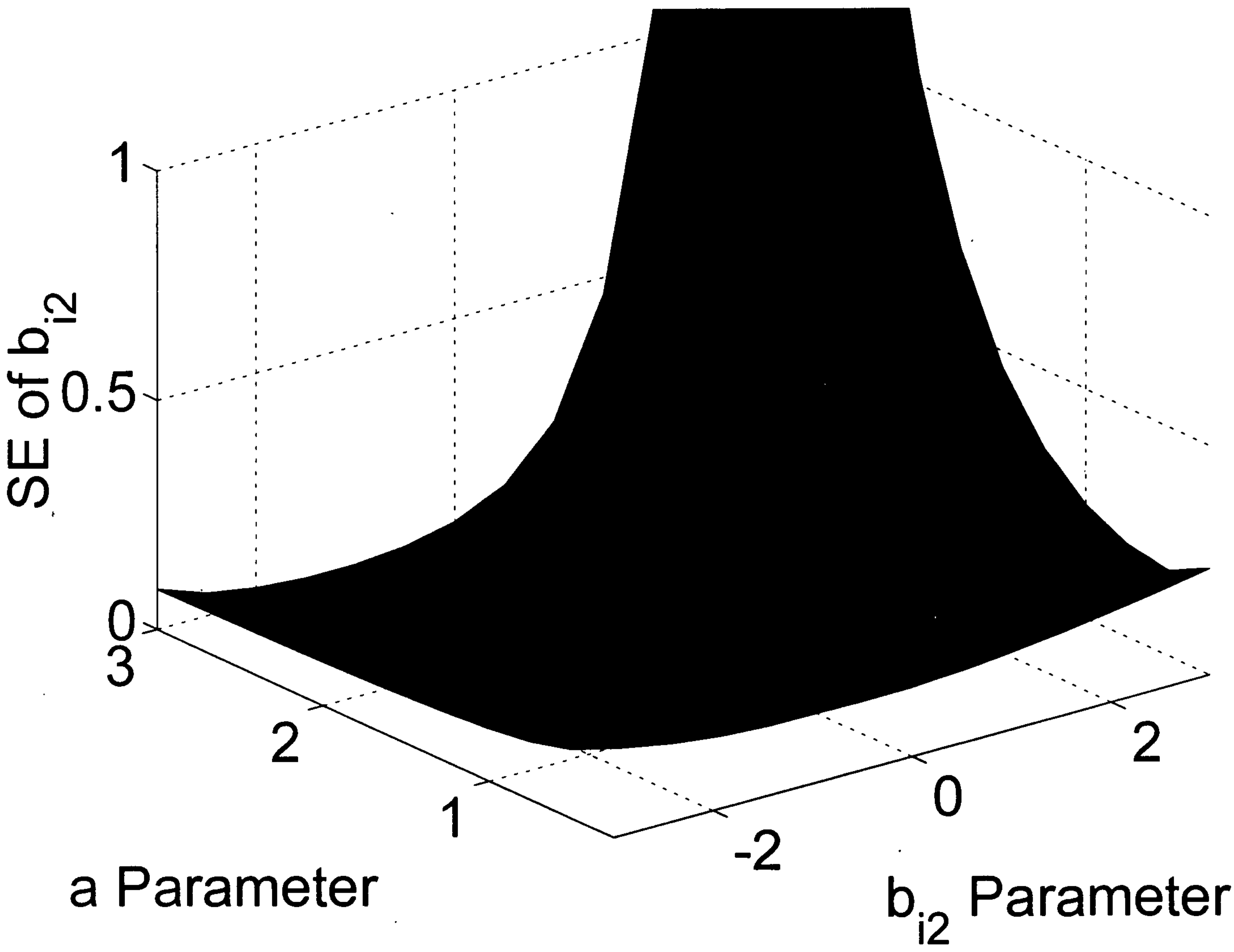


Figure 4a. The standard errors of item-category difficulties (b_{i2}) shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are -1 and 0.

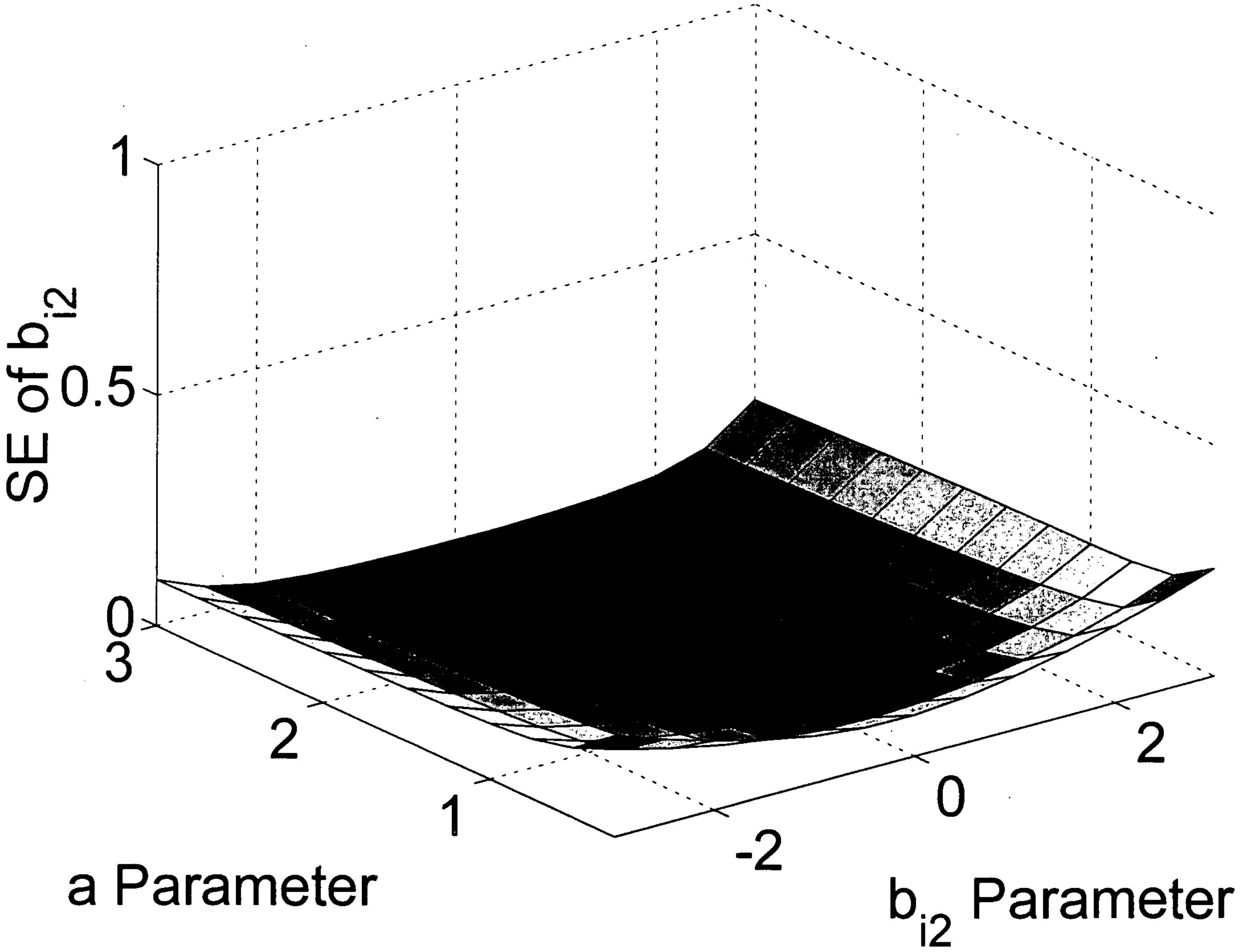


Figure 4b. The standard errors of item-category difficulties (b_{i2}) shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are 3 and 4.

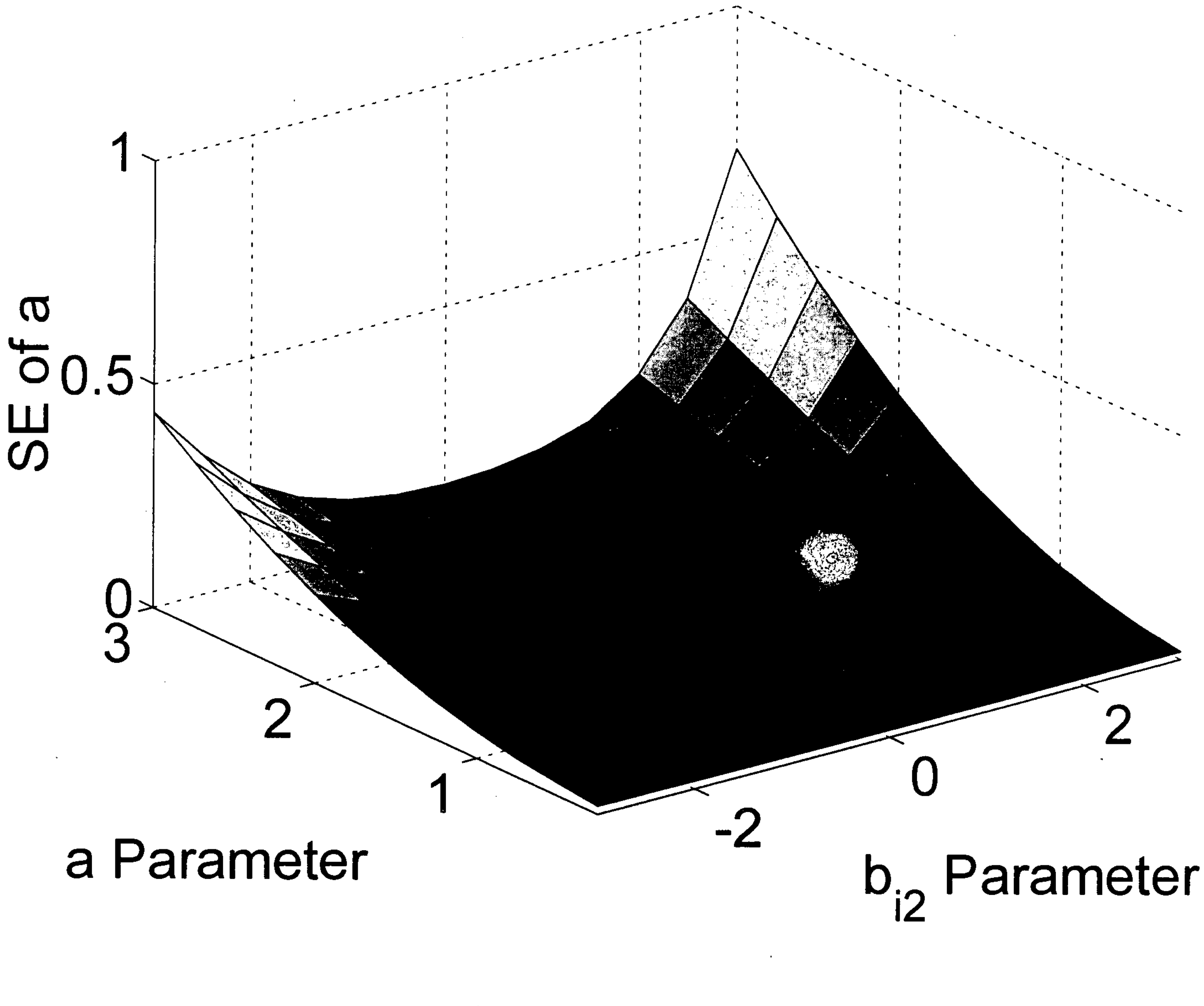


Figure 4c. The standard errors of item discriminations shown as the bivariate function of both tem-category difficulty (b_{i2}) and discrimination estimates for the GPCM model when the item-category difficulties, b_{i3} and b_{i4}, are 3 and 4.

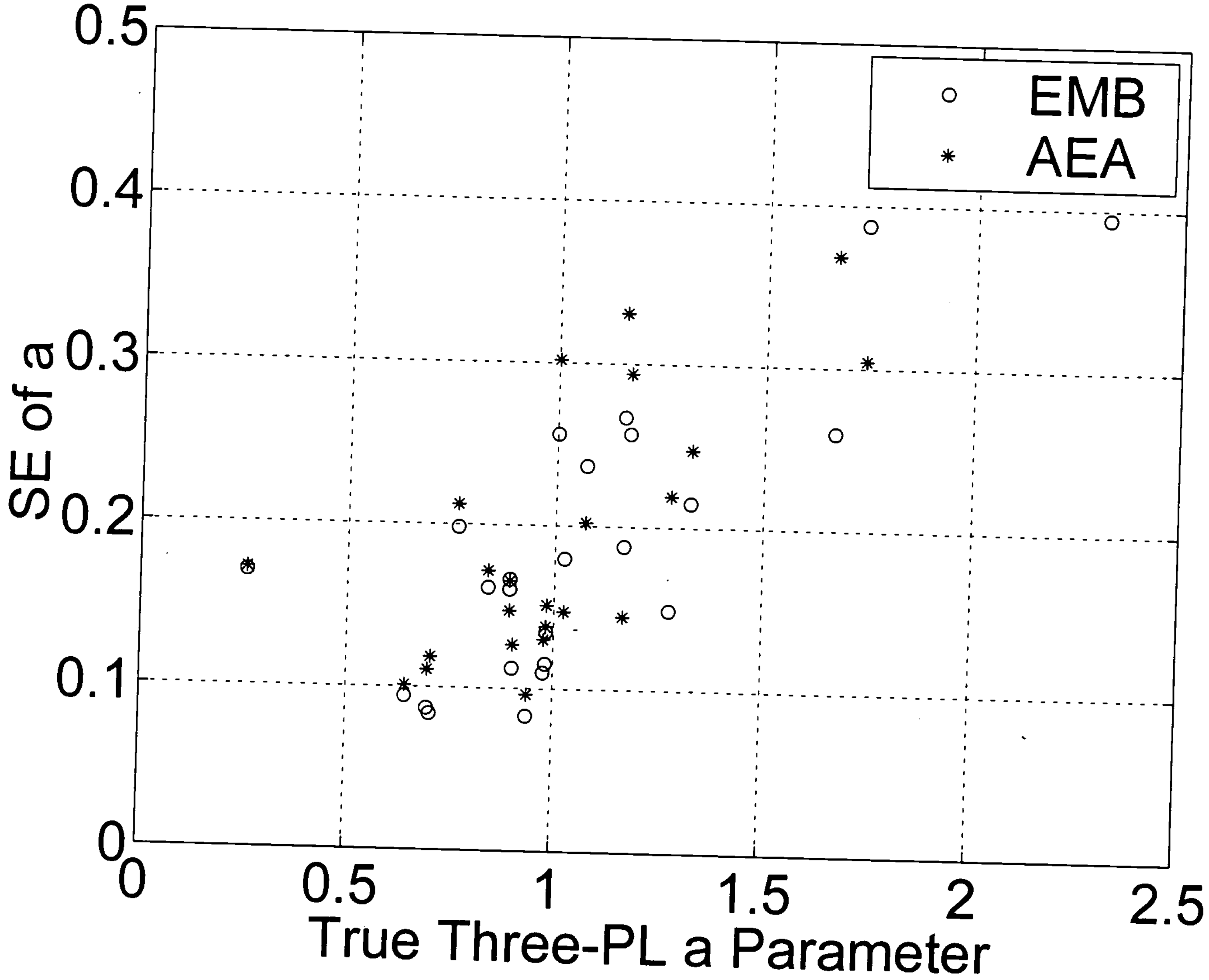


Figure 5a. SE of a as a function of the true a-parameter for the Three-PL model.

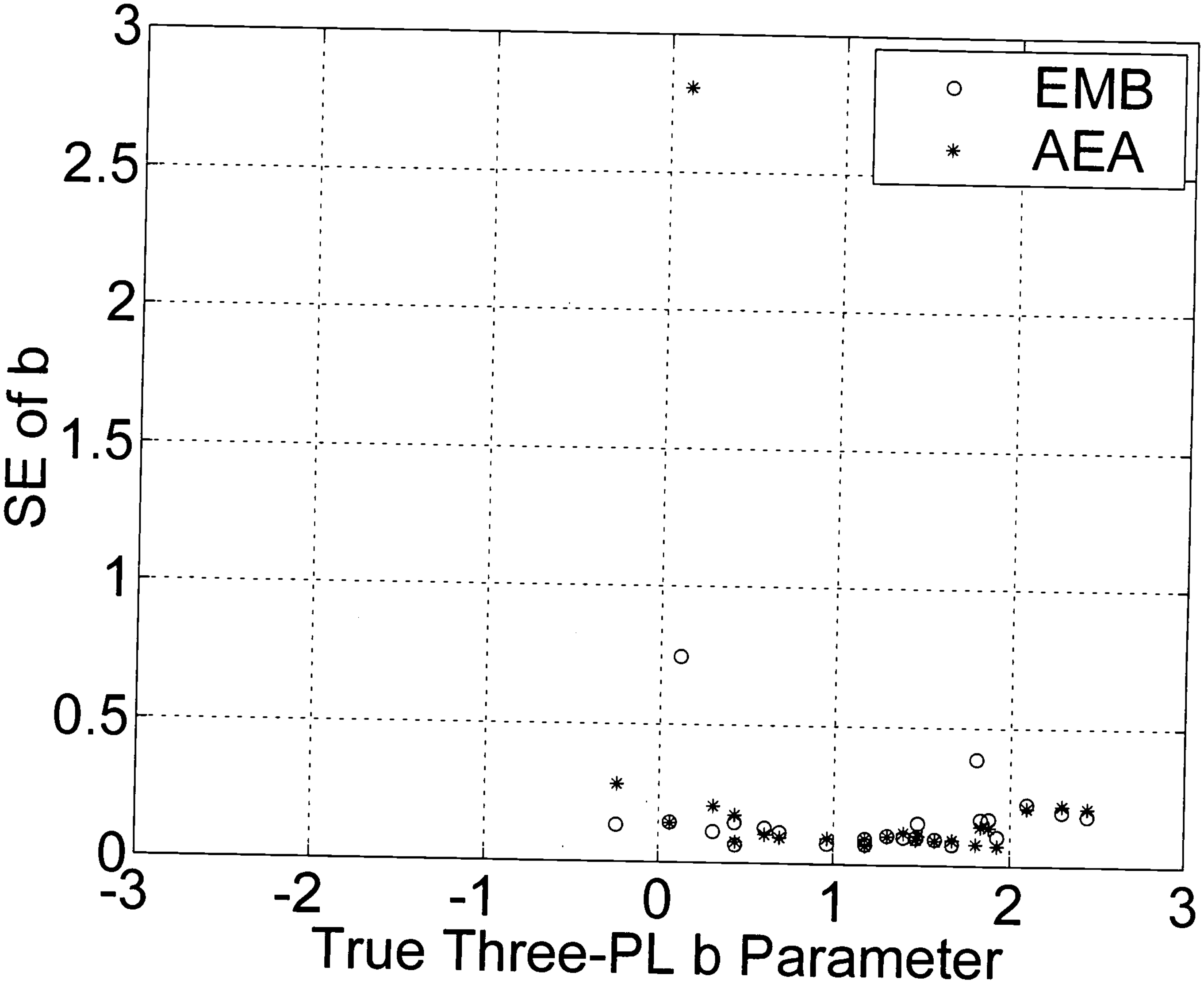


Figure 5b. SE of b as a function of the true b-parameter for the Three-PL model.

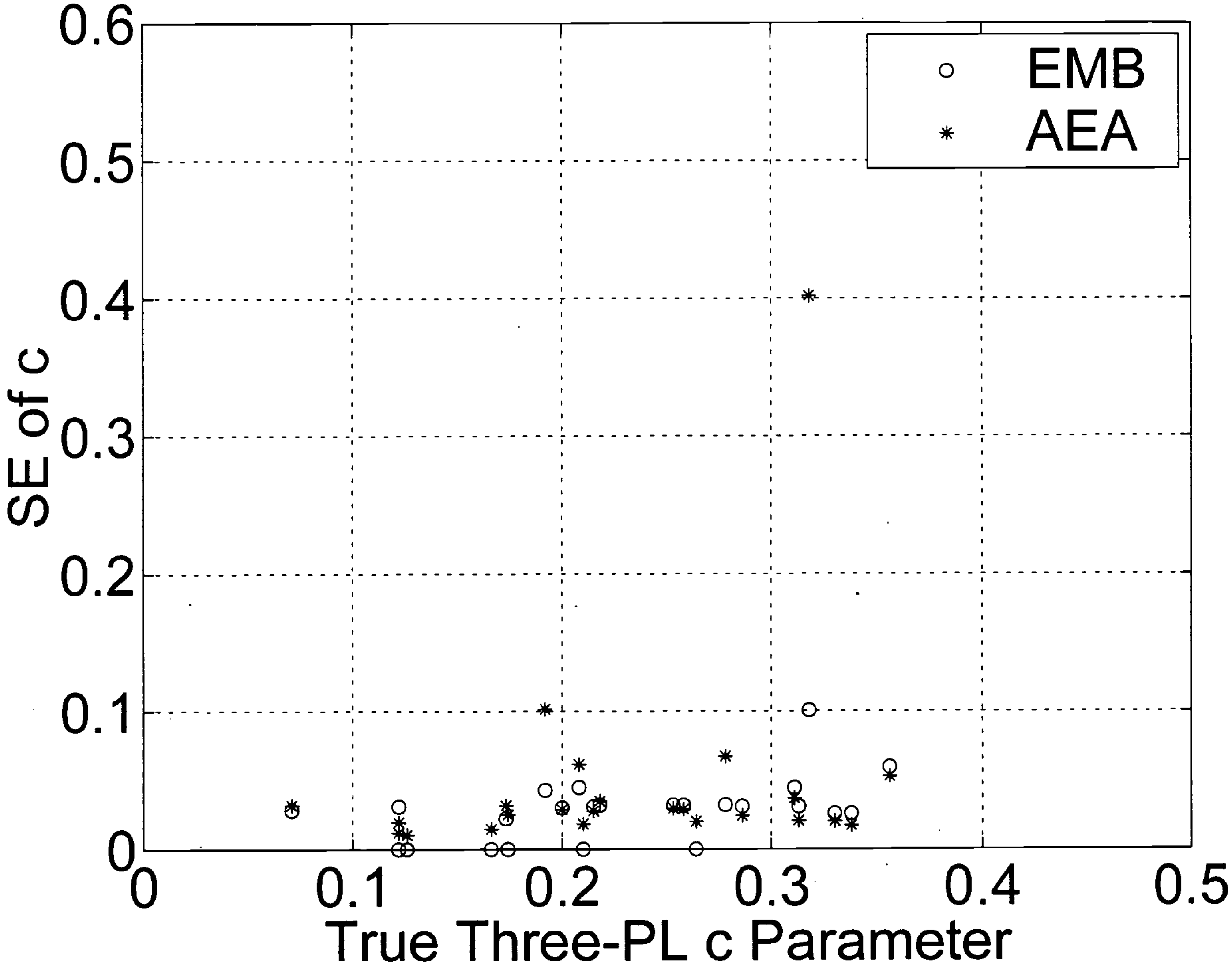


Figure 5c. SE of c as a function of the true c-parameter for the Three-PL model.

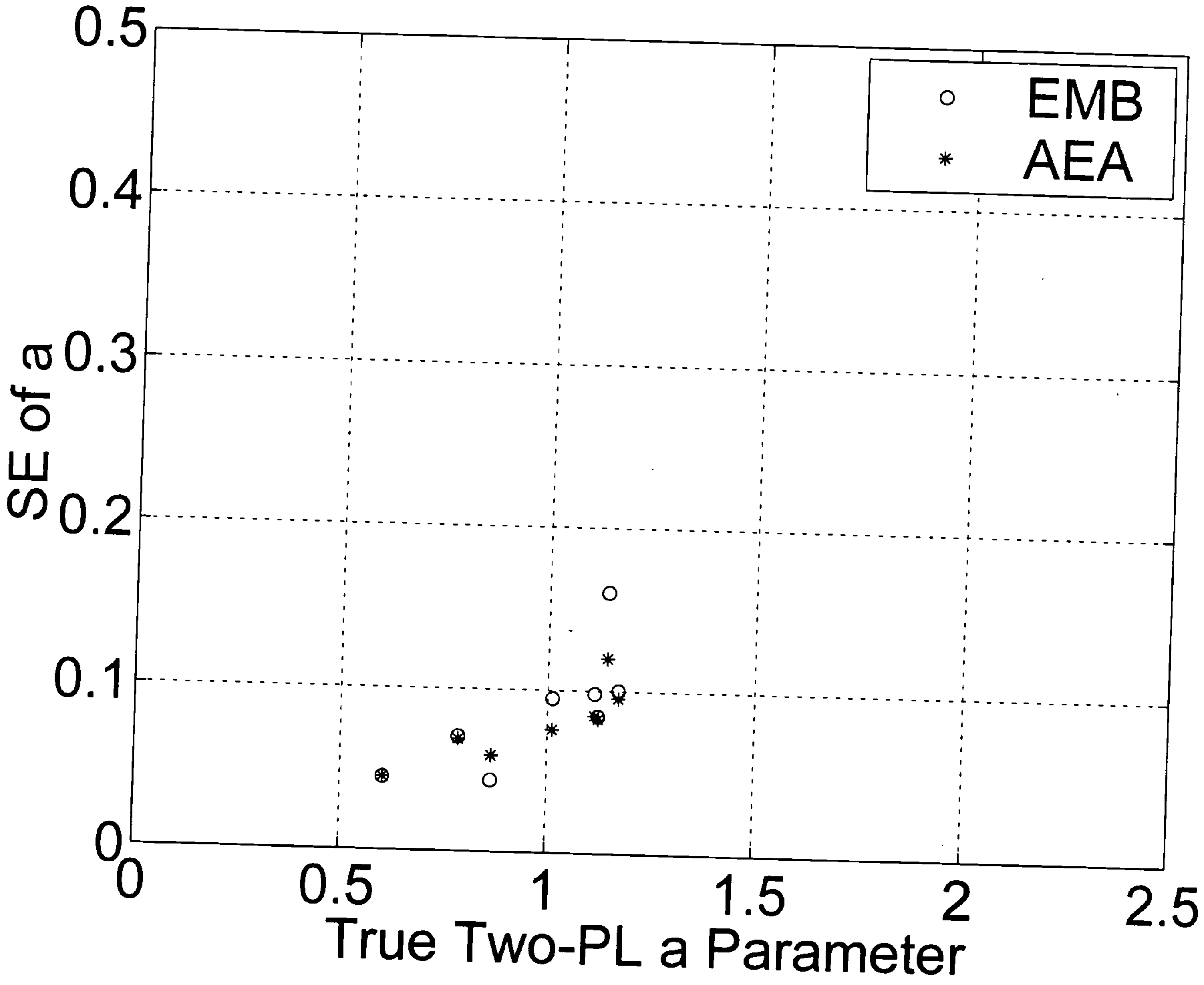


Figure 6a. SE of a as a function of the true a-parameter for the Two-PL model.

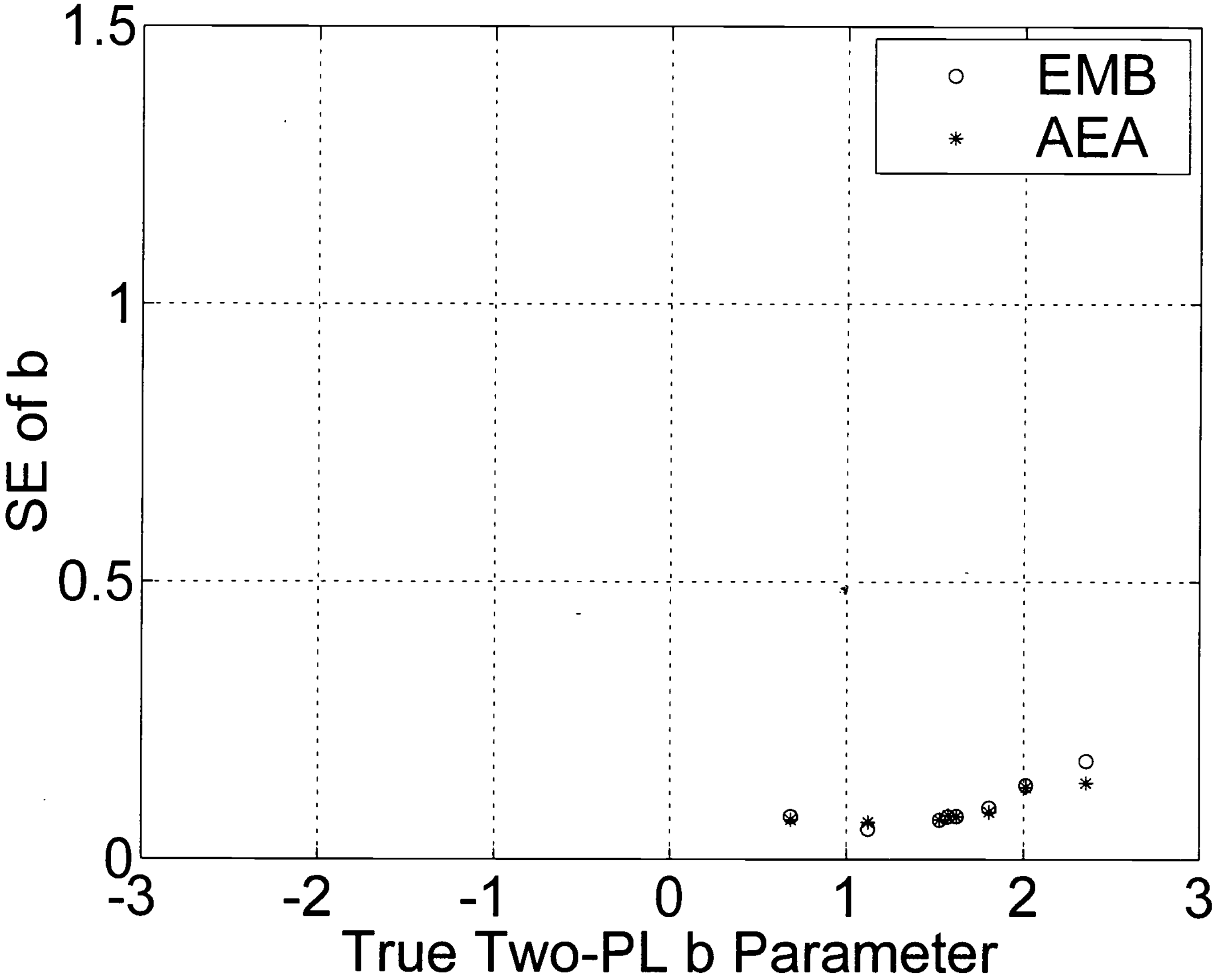


Figure 6b. SE of b as a function of the true b-parameter for the Two-PL model.

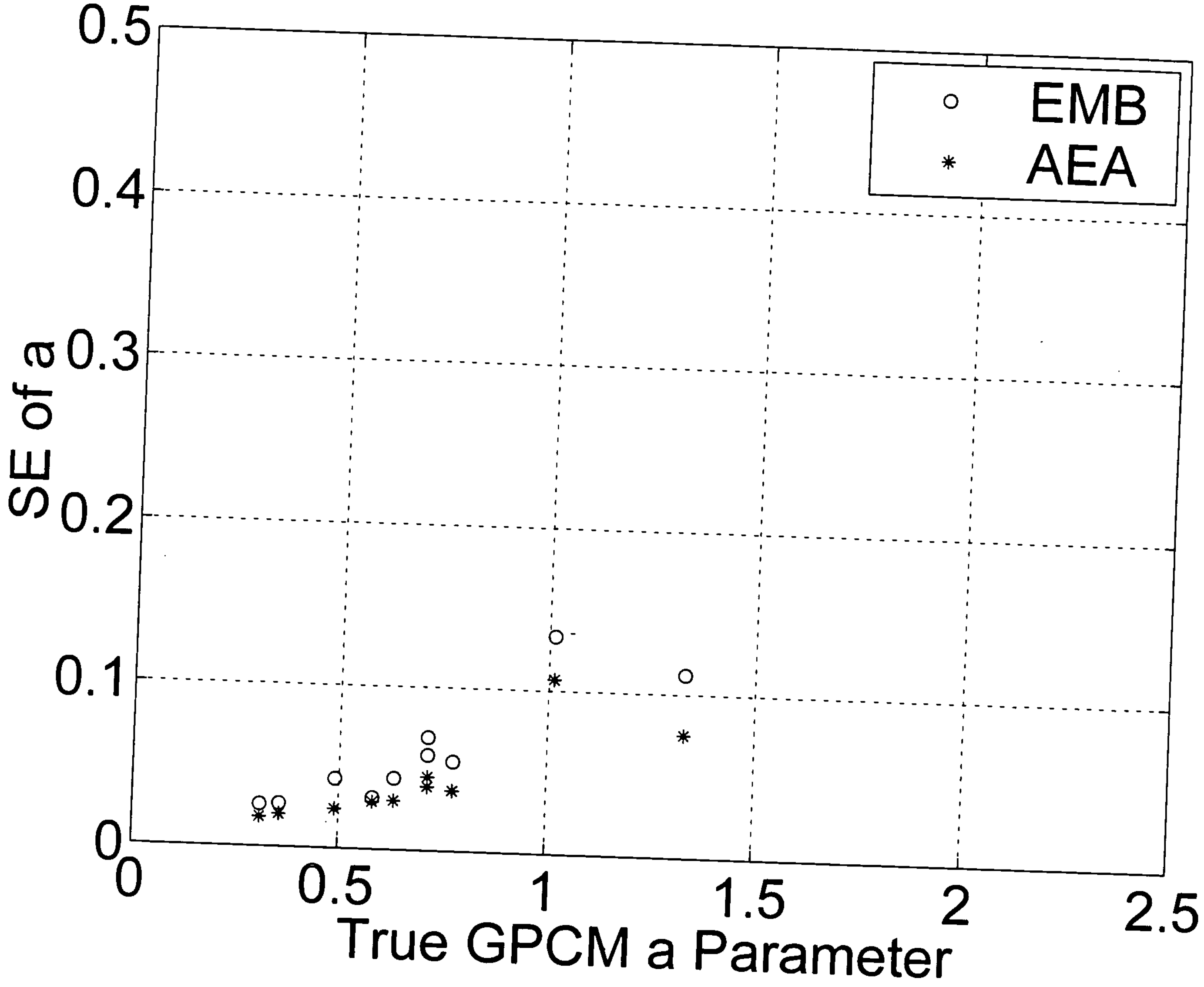


Figure 7a. SE of a as a function of the true a-parameter for the GPCM model.

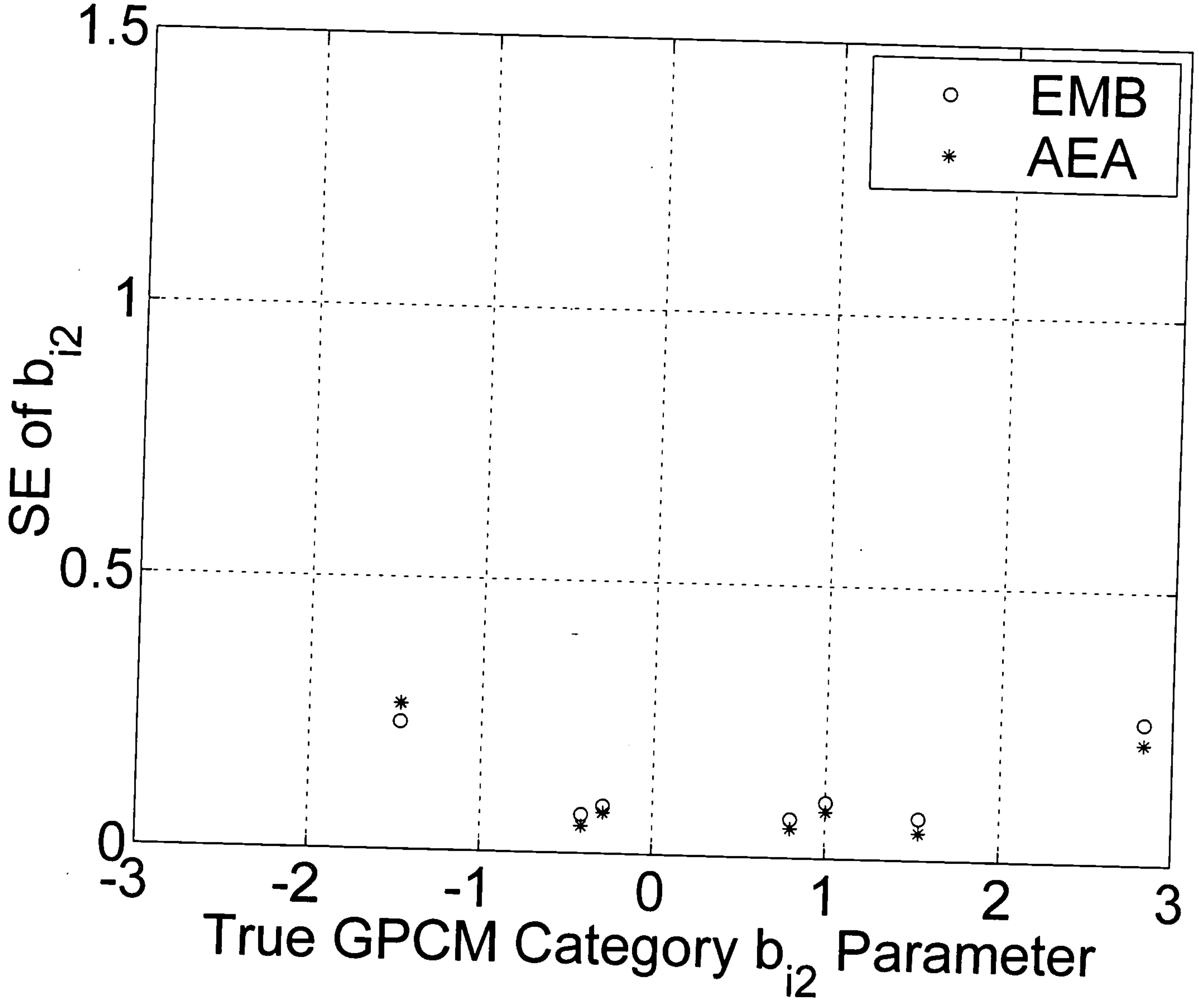


Figure 7b. SE of b_{i2} as a function of the true item-category parameter, b_{i2}, for the GPCM model.

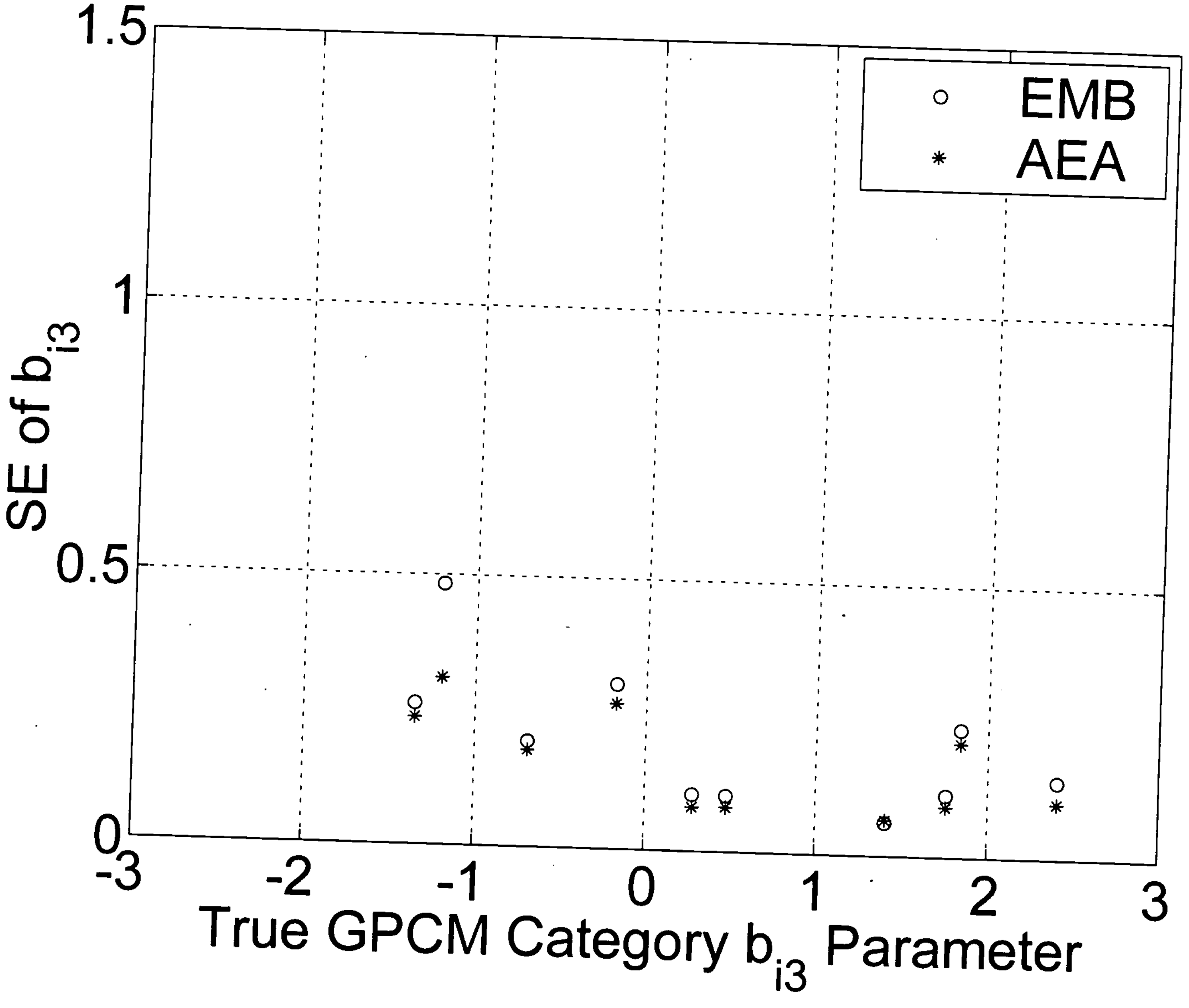


Figure 7c. SE of b_{i3} as a function of the true item-category parameter, b_{i3}, for the GPCM model.

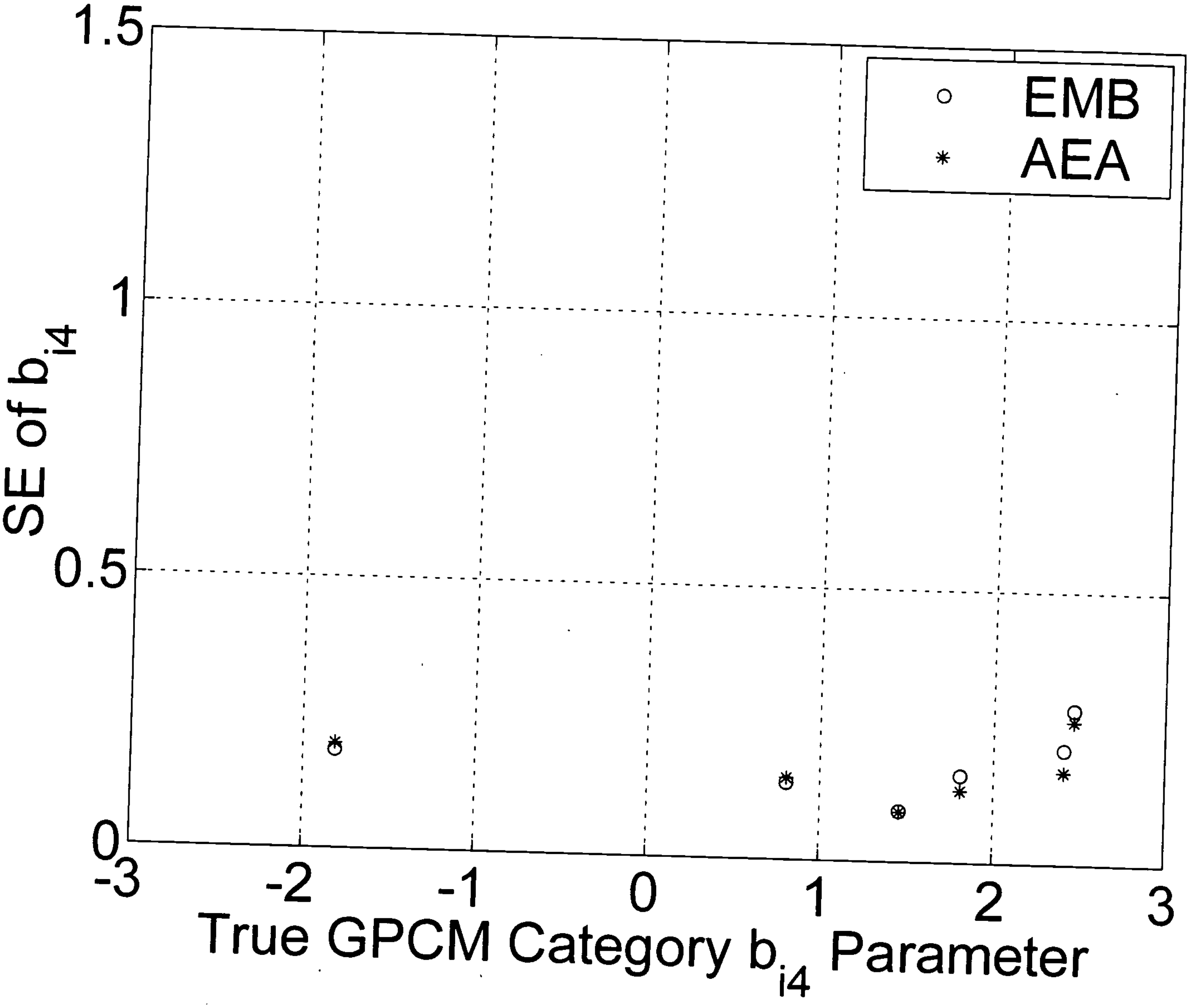


Figure 7d. SE of b_{i4} as a function of the true item-category parameter, b_{i4}, for the GPCM model.

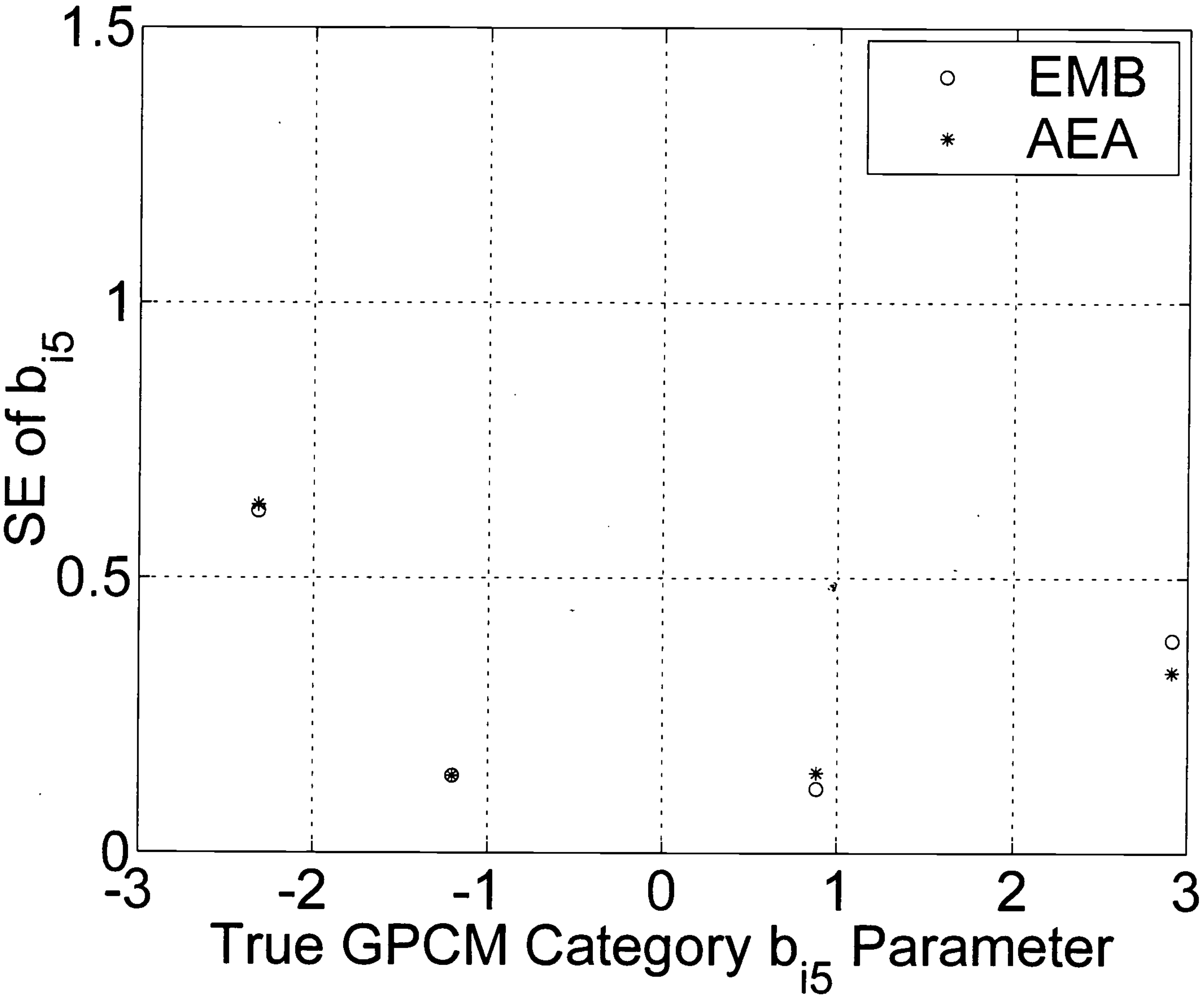


Figure 7e. SE of b_{i5} as a function of the true item-category parameter, b_{i5}, for the GPCM model.

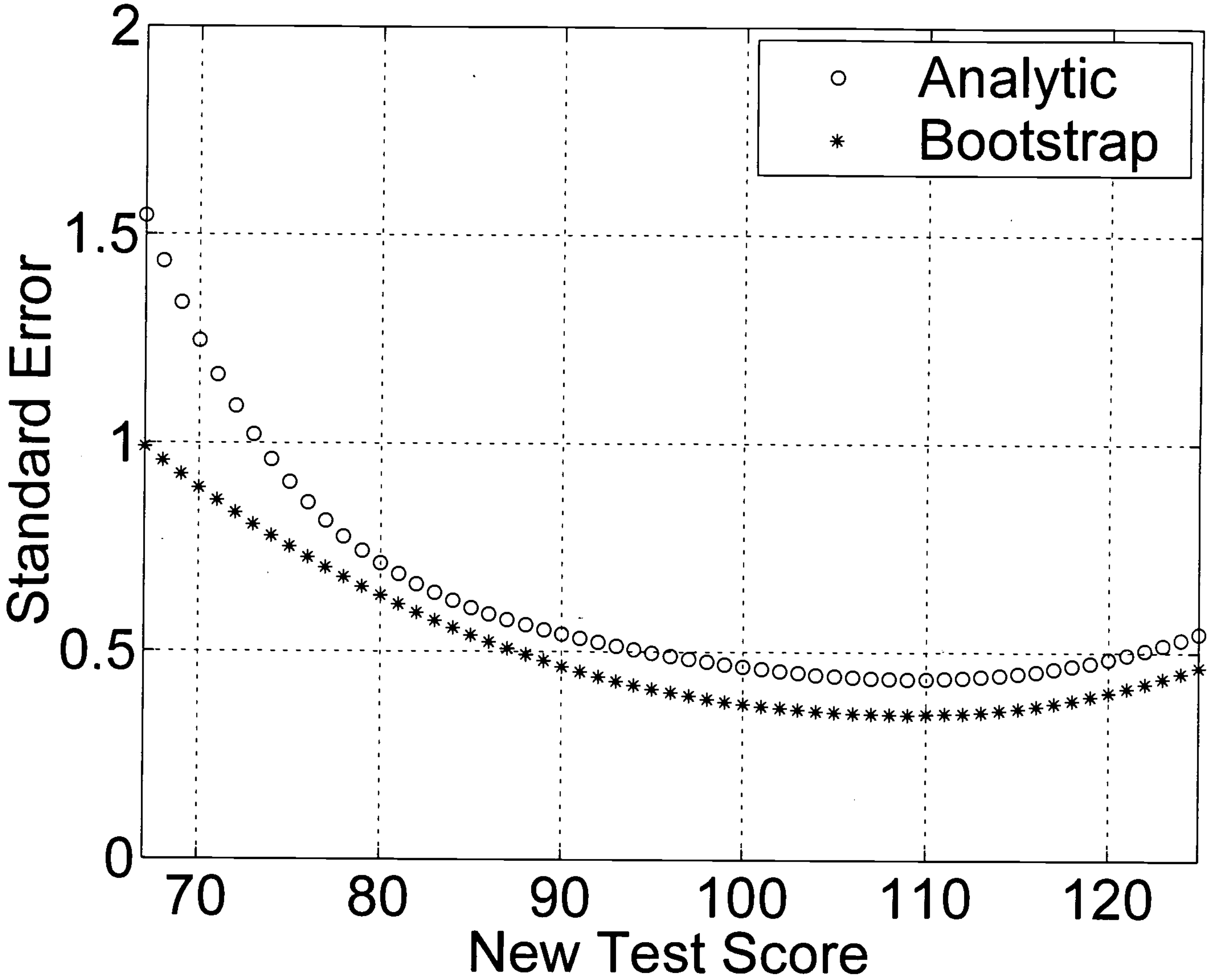


Figure 8. Standard Error as a Function of Raw Score for the Analytic and Bootstrap Method

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031641

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Applications of the Analytically Derived Asymptotic Standard Errors of IRT Item Parameter Estimates

Author(s): Yuan H. Li & Robert W. Lissitz

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→

Signature: Yuan H Li

Printed Name/Position/Title: Yuan H. Li / Statistical Specialist

elephone: 1-952-6764

FAX: 301-952-6228

Mail Address:

Date: 7/13/00

Name: Yuan H. LI
ss: Prince George's County Public Schools
Room 205
Upper Marlboro, MD. 20772

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.